Audiovisual Summarization of Lectures and Meetings Using a Segment Similarity Graph

Chidansh Bhatt*
FX Palo Alto Laboratory, Inc.
3174 Porter Drive
Palo Alto, CA, USA 94304
bhatt@fxpal.com

Andrei Popescu-Belis
Idiap Research Institute
Rue Marconi 19, CP 592
Martigny, Switzerland 1920
apbelis@idiap.ch

Matthew Cooper
FX Palo Alto Laboratory, Inc.
3174 Porter Drive
Palo Alto, CA, USA 94304
cooper@fxpal.com

ABSTRACT

We propose a method for extractive summarization of audiovisual recordings focusing on topic-level segments. We first build a content similarity graph between all segments across the collection, using word vectors from the transcripts, and then select the most central segments for the summaries. We evaluate the method quantitatively on the AMI Meeting Corpus using gold standard reference summaries and the ROUGE metric, and qualitatively on lecture recordings using a novel two-tiered approach with human judges. The results show that our method compares favorably with others in terms of ROUGE, and outperforms the baselines for human scores, thus also validating our evaluation protocol.

Keywords

Multimedia summarization, task-based evaluation, content-based similarity

1. INTRODUCTION

Audiovisual summarization is a key technique for efficient, large scale video retrieval. It uses structural characterizations of multimedia to generate proxies for longer streams that represent the essence of the original content with minimal redundancy. In this paper, we propose an approach to generate summaries of lectures or meetings from large repositories. We use speech transcripts and topic-level segments, and design a graph-based method, which builds a content-based segment similarity graph and selects the most connected segments for inclusion in summaries. A first evaluation using ground-truth extractive summaries from the AMI Meeting Corpus shows that our method outperforms two state-of-the-art methods. A second evaluation uses an innovative two-tiered protocol which relies on subjects writing summaries based on the automatic audiovisual summaries of lectures, and human judges rating their written summaries only. The results show that our method outperforms two baselines, and validate the protocol.

*Work performed at the Idiap Research Institute and at the FX Palo Alto Laboratory, Inc.

ICMR'16, June 06-09, 2016, New York, NY, USA

DOI: http://dx.doi.org/10.1145/2911996.2912047

2. RELATED WORK

Existing methods for A/V summarization can be categorized based on the modalities they use: text, audio, video, or a combination. Surveys are available for individual modalities, e.g. for text [14] or video summarization [13]. The style and extent to which various modalities are used within summaries varies greatly. Here we focus on extractive video summarization based on text speech transcripts, leveraging automatically detected topic structure.

To generate extractive summaries based on speech transcripts, several methods exist to rank sentences according to relevance metrics. In supervised approaches, a classifier is trained on sentence features such as keywords, sentence length, and position. Unsupervised approaches like Maximal Marginal Relevance [17] typically create a term vector per sentence.

Graph-based methods include eigenvector centrality approaches such as TextRank [12] and LexRank [5]. Extensions to such approaches include ClusterRank [6], which segments the transcript by subject. Like ClusterRank, our approach relies on topic-coherent segments; however, excerpts are selected with a different criterion.

3. TopS SUMMARIZATION METHOD

The proposed summarization method, named *TopS*, proceeds in three stages: topic segmentation of the ASR transcripts, computation of a content similarity graph relating all segments of the collection, and selection of the most connected segments as the summary of each recording. Finally, to create viewable audiovisual summaries, we sort the segments by order of occurrence in the recording, and append them together separated by fading effects for smooth transitions, to enhance the viewing experience.

3.1 Topic Segmentation

Most previous work on extractive summarization has focused on ranking *sentences*, but here, to avoid separating sentences from their contexts, we obtain *topic-level segments* of each lecture or meeting, using transcripts from automatic speech recognition (ASR) [7]. We use TextTiling [8] as implemented in the NLTK toolkit [2], for its robustness and simplicity. In TextTiling, topic shifts are determined from lexical co-occurrence patterns, computed here from 20-word

pseudo-sentences. Topic similarity scores are computed at sentence gaps using block comparison. Boundaries are set at peak differences between scores, adjusted to the closest speech segment breaks.

3.2 Centrality-based Summarization

We create a word-vector representation of each segment, after conversion to lower-case, tokenization, and stop word removal. We consider only unigrams, with a vocabulary size of 20,000 words. Each topic-level segment is indexed in the word-vector space with tf-idf weights for its words. Given the corpus C with a total of M segments, we consider the segment-indexed feature vectors $S = \{\mathbf{s}_i : i = 1, \ldots, M\}$. The similarity between segments $\mathbf{s_i}$ and $\mathbf{s_j}$ is computed as the cosine similarity between their word-vectors. We thus generate an M-by-M segment similarity score matrix, where each row corresponds to a segment and the cells of the row represent its similarity scores to the other ones. The matrix can be seen as a segment similarity graph, with a node for each segment and edges represented by the matrix entries and can be used for recommendation or hyperlinking tasks as well [1].

Graph centrality is an established criterion for extractive summarization [15], typically at the sentence level. Intuitively, the central segments of the graph, i.e. those that are strongly related to other segments in the collection (multiple document relations) or within the document (single document relations), should be preferred for inclusion in a summary of the lecture or meeting from which they originate. We propose a summarizer that maximizes centrality relative to either a single document or the entire collection.

We compute a global relevance score for each segment $\mathbf{s_i}$ as the sum of its similarity scores with all other segments in the graph: $\mathit{Global}(s_i) = 1/M \sum_{j=1}^{M} \mathrm{sim}(\mathbf{s_i}, \mathbf{s_j})$. Similarly, a local relevance score for each segment $\mathbf{s_i}$ is computed as the summation of its similarity scores over the nodes in the sub-graph of its document, divided into segments s_p to s_r as explained in Section 3.1, as $\mathit{Local}(s_i) = 1/(r-p) \sum_{j=p}^{r} \mathrm{sim}(\mathbf{s_i}, \mathbf{s_j})$.

To build the summary of a recording, its segments are sorted by decreasing relevance scores (either *Global* or *Local*), and the top scoring ones are added to the summary, until the desired duration is reached. Typically, this duration is a proportion of the original recording. For instance, as indicated by Mani et al. [11] in the case of written material, summaries as short as 17% of the full text's length speed up decision-making by a factor of two.

Unlike previous graph-based approaches to summarization, our method does not consider sentences but larger topic-level segments, to improve the local coherence of the summary. Moreover, while many graph-based approaches ignore edges with very low weights, we use all edges to determine the most important segments within a recording.

3.3 Hybrid Summarization

Hybrid approaches combine the *Local* and *Global* relevance scores and domain-specific features. For instance, we consider the sum (or product) of *Local* and *Global* to enforce both properties during segment selection. Or, we can consider *Local–Global* (or even *Local*/*Global*) to select segments having high similarity within the document and low similarity within the corpus as in discriminative key-frame selection approach [4].

Table 1: Scores of *TopS* and two other methods.

	ROUGE 1			ROUGE 2		
Method	P	R	F	P	R	F
TopS(Global)	**.66**	**.70**	**.65**	**.42**	**.43**	**.41**
TextRank	.30	.21	.24	.05	.03	.04
ClusterRank	.32	.26	.28	.05	.04	.04

TopS can be combined with genre-specific features for multi-party meetings. The first one is the *number of unique speakers* in each topic segment, e.g., a speaker sequence of A-A-A-C-D has 3 unique speakers. The second one is the *number of speaker changes* (or turns) in each topic segment, e.g., 2 in the example above. We hypothesize that segments with more unique speakers and with more speaker changes are more active and thus more relevant to a summary. These features can be used with the *Global* and *Local* scores in *TopS*, noted as *TopS*(Global + Change + Unique).

4. REFERENCE-BASED EVALUATION

For evaluation, we use all available reference extractive summaries for meetings of the AMI Meeting Corpus [3]. Each meting has an ASR transcript of 3000–7000 words, and topic segmentation generates an average of 50 segments per meeting. The reference summaries contain utterances selected by human judges as best representing each meeting's content. We measure the overlap between these reference summaries and those found by *TopS*(Global) using ROUGE [10], a metric that counts the number of overlapping n-grams (up to $n = 4$). ROUGE-1 was argued to be the most relevant version for meeting summarization [16].

The length of the *TopS* summaries is set at 17% of each meeting's length. We compare the ROUGE scores of *TopS* with those obtained by TextRank and ClusterRank, using respectively the scores from [12] and [6].

Table 1 shows that *TopS* (Global) outperforms TextRank and ClusterRank for both levels of ROUGE (n=1 and n=2) for which TextRank and ClusterRank scores are available, with more than 110% improvement in terms of average score. Another approach using lexical and prosodic features [9] achieves a ROUGE-1 F-score of 0.59, still below the *TopS* F-score of 0.65. As for n=3 and n=4, ROUGE F-scores of *TopS*(Global) are 0.26 and 0.22 respectively, with slightly higher precision scores (0.30 and 0.26 respectively).

Fig. 1 represents the performance of several variants of our approach. *TopS* using *Global* similarity marginally outperforms *Local* similarity (and hence their sum), a result related to the 80% overlap between the summary segments of *Global* and *Local*. The use of *Local–Global* similarity to create summaries with more locally unique segments leads to higher precision but lower recall and F1 scores. The domain-specific features (Speaker Change and Unique Speaker) score below the graph-based ones, especially in terms of recall, and their combination with *TopS* does not improve performance. In sum, we conclude that *TopS* summarization outperforms other existing methods in our evaluation.

5. SUBJECTIVE EVALUATION

While there is a growing consensus on evaluating meeting summarization, other genres lack widely accepted evaluation protocols. We now describe an efficient and broadly applicable approach to evaluating audiovisual summaries of

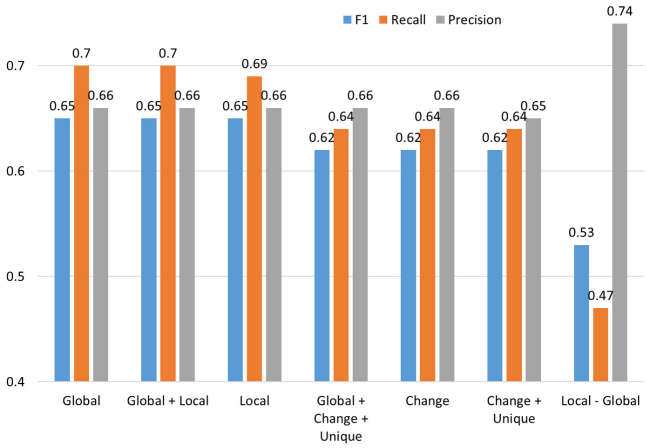

Figure 1: Performance of TopS and hybrid combinations according to the ROUGE-1 metric.

recordings of university lectures and scientific conferences, and apply it to the *TopS* method.

5.1 Evaluation Protocol

We consider 169 recordings with ASR transcripts lectures recorded and distributed by Klewel (www.klewel.com). As human-generated extractive summaries are not available to compute ROUGE, we propose a two-tiered evaluation protocol involving human subjects. This task-based protocol avoids requiring judges to rate summaries by comparison with the entire recording, or to define reference ones.

The task we have in mind is the production by a human of a *written summary* of a recording after watching the *automatically created audiovisual summary*. The written summaries are then rated by judges who view the entire recording—but only once for potentially multiple summaries, methods and/or human subjects. The judges can thus easily compare multiple summaries, and we propose two dimensions for this comparison, to be rated on 5-point scales: *fidelity*, i.e. how much of the original content is rendered in the summary, and *fluency*, i.e. how correct is the English used in the summary. These are distinguished in order not to penalize subjects who render correctly the content of lecture but do not write English fluently. The evaluation protocol consists of the following steps:

1. Subjects watch an A/V summary online, possibly multiple times with play/pause/rewind.
2. Subjects write summaries of the lectures, using their own words as much as possible.
3. Judges watch the original A/V recording and rate one or more summaries in terms of fidelity and fluency.
4. The ratings are averaged over recordings, subjects and judges into one score for each summarization method.

Our hypothesis is that better summarization methods lead to better written summaries and higher scores. This evaluation protocol is efficient because subjects (writers) only have to watch short A/V summaries, while judges, who must view an entire recording, can then rate an arbitrary number of written summaries. Subjects must not watch different summaries of the same recording, to avoid learning effects. Such effects are a potential drawback of traditional subjective evaluation of summaries, in which each judge watches the original recording as well as different summaries before they provide their judgment. In the experiments below, we show that this evaluation method can accommodate variability across recordings, subjects, and judges to provide conclusive results.

We conducted two evaluations based on this protocol: a pilot and a full study. Their main goal is to compare *TopS* with two common baseline methods: either display an excerpt from the very beginning of a recording (a method noted as *First*), or select several random segments (noted *Random*). As we use topical segments from TextTiling, this likely improves the coherence of *Random* compared to using uniform segments. For all methods, summaries are again 17% of the ASR transcript's length.

5.2 Pilot Study

In the pilot study, we selected three video lectures from Klewel and generated A/V summaries for each of them using the three methods above. We showed to three subjects one summary of each video, each from a different method (3x3 Latin square design), instructing them to write a summary of at most 100 words. The nine resulting summaries were rated by two judges each. On average, users spent 10 minutes to watch each A/V summary and to write their textual summary. The average summary length was 72 words. *TopS* resulted in the highest fidelity scores from both judges, while *Random* was second and *First* was third.

5.3 Full Study

In the full study, we selected six lecture recordings from Klewel (not used in the pilot study) and generated A/V summaries using the same automatic methods: *First*, *Random* and *TopS*. We showed to each of the six subjects (English-speaking PhD students at our institution who were paid for their participation) one summary of each video, distributed evenly over the methods—hence each subject saw two summaries from each method, without knowledge of the method, but never more than one summary of the same video. Conversely, each of the three summaries of each video was seen by two subjects, to measure agreement. Finally, all 36 written summaries were rated by three different judges (English-speaking researchers in the field) who saw the videos in full before rating the written summaries.

Subjects spent on average 12 minutes to watch each A/V summary and write their textual summary, again a reasonable duration given that lectures lasted between 30 and 60 minutes. The average summary length was 84 words.

The fidelity scores for each summarization method when averaging over lectures and judges (but not subjects) appear in Table 2. For 5 of the 6 subjects, the fidelity is highest for *TopS*. The differences between *TopS* and, respectively, *Random* and *First* represent consistent trends at the 90% level (paired t-test across subjects, $p < 0.1$), while the difference between *Random* and *First* is not significant.

Table 3 shows the fidelity scores per method. This time we keep the lectures separate, as opposed to the different subjects, and we average over the two subjects seeing each A/V summary and the three judges rating the written summaries. The results indicate that for four lectures out of six (lectures 3 to 6), fidelity is highest for *TopS*. On average, *TopS* significantly outperforms *Random* (paired t-test across subjects, $p < 0.01$), though not *First*. Still, *TopS* also outperforms *First* if we exclude the first two lectures, which

Table 2: Fidelity of human written summaries based on three A/V summarization methods, for six subjects, *averaged over lectures and judges.*

Subject	*First*	*Random*	*TopS*
1	3.00	3.17	**4.33**
2	2.83	2.50	**3.67**
3	3.17	**3.75**	2.75
4	3.17	1.75	**3.33**
5	2.08	1.75	**2.58**
6	2.58	1.92	**3.50**
Average	2.81	2.47	**3.36**

happen to start with an overview that actually provides a summary, which explains the good results of *First*. These two lectures are courses in software engineering, while the other four are short non-technical conference presentations.

Table 3: Fidelity of human written summaries based on three different A/V summarization methods, for six lectures, *averaged over subjects and judges.*

Lecture	*First*	*Random*	*TopS*
1	**2.92**	2.25	2.58
2	**3.33**	2.83	3.25
3	1.67	3.08	**3.92**
4	2.83	1.42	**3.08**
5	2.83	2.67	**3.67**
6	3.25	2.58	**3.67**
Average	2.81	2.47	**3.36**

Although less marked, there are differences in fluency as well. The average fluency per method is 4.21 vs. 4.06 vs. 3.82 respectively for *TopS*, *First*, and *Random*. *TopS* significantly outperforms *First* ($p < 0.05$) but not *Random* (despite the larger difference). As expected, the average fluency per subject varies quite a lot (the values are: 4.72, 4.25, 4.28, 3.31, 3.47, and 4.14). However, when looking at fluency per lecture, no significant differences appear across the three methods, as the values are closer because they are averaged over the same subjects.

6. CONCLUSION

We have presented *TopS*, a graph-based method for extractive summarization of audiovisual lecture recordings. *TopS* leverages topic segmentation and content-based similarity at the collection level to provide a simple yet effective approach to summarization. *TopS* is computationally efficient and compares favorably with state-of-the-art methods. Our experiments validate *TopS* using gold-standard summaries from the AMI Meeting Corpus and the ROUGE metric, but also, when no reference summary was available, using a task-based approach which considers audiovisual extractive summaries as a means for producing human-written summaries of lectures. In the future, we aim to enrich content-similarity with visual features and extend the content-similarity graph using collaborative filtering and per-segment view count information, in order to integrate user behavior and preferences.

7. ACKNOWLEDGMENTS

We are grateful to the Swiss National Science Foundation for its support (AROLES project n. 51NF40-144627).

8. REFERENCES

[1] C. Bhatt, N. Pappas, M. Habibi, and A. Popescu-Belis. Multimodal reranking of content-based recommendations for hyperlinking video snippets. In *Proc. of Int. Conf. on Multimedia Retrieval (ICMR)*, pages 225–232, 2014.

[2] S. Bird. NLTK: the Natural Language Toolkit. In *Proc. of the COLING/ACL Interactive Presentation Sessions*, 2006.

[3] J. Carletta. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Lang. Resources & Evaluation*, 41(2):181–190, 2007.

[4] M. Cooper and J. Foote. Discriminative techniques for keyframe selection. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2005.

[5] G. Erkan and D. R. Radev. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 2004.

[6] N. Garg, B. Favre, et al. ClusterRank: a graph based method for meeting summarization. In *Proc. INTERSPEECH*, 2009.

[7] T. Hain et al. Transcribing meetings with the AMIDA systems. *IEEE Trans. on Audio, Speech, and Language Processing*, 20(2):486–498, 2012.

[8] M. A. Hearst. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

[9] C. Lai and S. Renals. Incorporating lexical and prosodic information at different levels for meeting summarization. In *Proc. of Interspeech*, 2014.

[10] C.-Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proc. of NAACL-HLT*, 2003.

[11] I. Mani, G. Klein, D. House, L. Hirschman, T. Firmin, and B. Sundheim. SUMMAC: A text summarization evaluation. *Natural Lang. Eng.*, 8(1):43–68, 2002.

[12] R. Mihalcea and P. Tarau. TextRank: bringing order into texts. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2004.

[13] A. G. Money and H. Agius. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation*, 19(2):121–143, 2008.

[14] A. Nenkova and K. McKeown. A survey of text summarization techniques. In *Mining Text Data*. 2012.

[15] R. Ribeiro and D. M. de Matos. Revisiting centrality-as-relevance: Support sets and similarity as geometric proximity. *Journal of Artificial Intelligence Research*, 42:275–308, 2011.

[16] K. Riedhammer, B. Favre, et al. Long story short–global unsupervised models for keyphrase based meeting summarization. *Speech Communication*, 2010.

[17] D. Wang, S. Zhu, T. Li, and Y. Gong. Comparative document summarization via discriminative sentence selection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 7(1):2, 2013.