Intra-Speaker Topic Modeling for Improved Multi-Party Meeting Summarization with Integrated Random Walk

Yun-Nung Chen and Florian Metze
School of Computer Science, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213-3891, USA
{yvchen,fmetze}@cs.cmu.edu

Abstract

This paper proposes an improved approach to extractive summarization of spoken multi-party interaction, in which integrated random walk is performed on a graph constructed on topical/ lexical relations. Each utterance is represented as a node of the graph, and the edges' weights are computed from the topical similarity between the utterances, evaluated using probabilistic latent semantic analysis (PLSA), and from word overlap. We model intra-speaker topics by partially sharing the topics from the same speaker in the graph. In this paper, we perform experiments on automatically and manually generated transcripts. For automatic transcripts, our results show that intra-speaker topic sharing and integrating topical/ lexical relations can help include the important utterances.

1 Introduction

Speech summarization is an active and important topic of research (Lee and Chen, 2005), because multimedia/ spoken documents are more difficult to browse than text or image content. While earlier work was focused primarily on broadcast news content, recent effort has been increasingly directed to new domains such as lectures (Glass et al., 2007; Chen et al., 2011) and multi-party interaction (Banerjee and Rudnicky, 2008; Liu and Liu, 2010). We describe experiments on multi-party interaction found in meeting recordings, performing extractive summarization (Liu et al., 2010) on transcripts generated by automatic speech recognition (ASR) and human annotators.

Graph-based methods for computing lexical centrality as importance to extract summaries (Erkan and Radev, 2004) have been investigated in the context of text summarization. Some works focus on maximizing coverage of summaries using the objective function (Gillick, 2011). Speech summarization carries intrinsic difficulties due to the presence of recognition errors, spontaneous speech effect, and lack of segmentation. A general approach has been found very successful (Furui et al., 2004), in which each utterance in the document d, $U = t_1 t_2 ... t_i ... t_n$, represented as a sequence of terms t_i, is given an importance score:

$$\begin{aligned} I(U,d) &= \frac{1}{n}\sum_{i=1}^{n}[\lambda_1 s(t_i,d) + \lambda_2 l(t_i) \\ &+ \lambda_3 c(t_i) + \lambda_4 g(t_i)] + \lambda_5 b(U), \end{aligned} \quad (1)$$

where $s(t_i,d)$, $l(t_i)$, $c(t_i)$, $g(t_i)$ are respectively some statistical measure (such as TF-IDF), linguistic measure (e.g., different part-of-speech tags are given different weights), confidence score and N-gram score for the term t_i, and $b(U)$ is calculated from the grammatical structure of the utterance U, and $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and λ_5 are weighting parameters. For each document, the utterances to be used in the summary are then selected based on this score.

In recent work, Chen (2011) proposed a graphical structure to rescore $I(U,d)$, which can model the topical coherence between utterances using random walk within documents. Similarly, we now use a graph-based approach to consider the importance of terms and the similarity between utterances, where topical and lexical similarity are integrated in the graph, so that utterances topically or lexically similar to more important utterances are given higher scores. Using topical similarity can compensate the negative effects of recognition errors on similarity evaluated on word overlap to some extent. In addition, this paper proposes an approach of modeling intra-speaker topics in the graph to improve meeting summarization (Garg et al., 2009) using information from multi-party interaction, which is not available in lectures or broadcast news.

2 Proposed Approach

We apply word stemming and noise utterance filtering for utterances in all meetings. Then we construct a graph to compute the importance of all utterances.

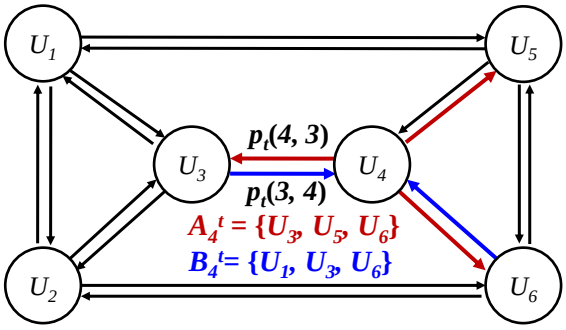

Figure 1: A simplified example of the graph considered.

We formulate the utterance selection problem as random walk on a directed graph, in which each utterance is a node and the edges between them are weighted by topical and lexical similarity. The basic idea is that an utterance similar to more important utterances should be more important (Chen et al., 2011). We formulate two types of directed edge, topical edges and lexical edges, which are weighted by topical and lexical similarity respectively. We then keep only the top N outgoing edges with the highest weights from each node, while consider incoming edges to each node for importance propagation in the graph. A simplified example for such a graph with topical edges is in Figure 1, in which A_i^t and B_i^t are the sets of neighbors of the node U_i connected respectively by outgoing and incoming topical edges.

2.1 Parameters from PLSA

Probabilistic latent semantic analysis (PLSA) (Hofmann, 1999) has been widely used to analyze the semantics of documents based on a set of latent topics. Given a set of documents $\{d_j, j = 1, 2, ..., J\}$ and all terms $\{t_i, i = 1, 2, ..., M\}$ they include, PLSA uses a set of latent topic variables, $\{T_k, k = 1, 2, ..., K\}$, to characterize the "term-document" co-occurrence relationships. The PLSA model can be optimized with EM algorithm by maximizing a likelihood function. We utilize two parameters from PLSA, latent topic significance (LTS) and latent topic entropy (LTE) (Kong and Lee, 2011) in the paper.

Latent Topic Significance (LTS) for a given term t_i with respect to a topic T_k can be defined as

$$\mathrm{LTS}_{t_i}(T_k) = \frac{\sum_{d_j \in D} n(t_i, d_j) P(T_k \mid d_j)}{\sum_{d_j \in D} n(t_i, d_j)[1 - P(T_k \mid d_j)]}, \quad (2)$$

where $n(t_i, d_j)$ is the occurrence count of term t_i in a document d_j. Thus, a higher $\mathrm{LTS}_{t_i}(T_k)$ indicates the term t_i is more significant for the latent topic T_k.

Latent Topic Entropy (LTE), for a given term t_i can be calculated from the topic distribution $P(T_k \mid t_i)$:

$$\mathrm{LTE}(t_i) = -\sum_{k=1}^{K} P(T_k \mid t_i) \log P(T_k \mid t_i), \quad (3)$$

where the topic distribution $P(T_k \mid t_i)$ can be estimated from PLSA. $\mathrm{LTE}(t_i)$ is a measure of how the term t_i is focused on a few topics, so a lower latent topic entropy implies the term carries more topical information.

2.2 Statistical Measures of a Term

The statistical measure of a term t_i, $s(t_i, d)$ in (1) can be defined in terms of $\mathrm{LTE}(t_i)$ in (3) as

$$s(t_i, d) = \frac{\gamma \cdot n(t_i, d)}{\mathrm{LTE}(t_i)}, \quad (4)$$

where γ is a scaling factor such that $0 \leq s(t_i, d) \leq 1$; the score $s(t_i, d)$ is inversely proportion to the latent topic entropy $\mathrm{LTE}(t_i)$. Some works (Kong and Lee, 2011) showed that the use in (1) of $s(t_i, d)$ as defined in (4) outperformed the very successful "significance score" (Furui et al., 2004) in speech summarization; then, we use it as the baseline.

2.3 Similarity between Utterances

Within a document d, we can first compute the probability that the topic T_k is addressed by an utterance U_i:

$$P(T_k \mid U_i) = \frac{\sum_{t \in U_i} n(t, U_i) P(T_k \mid t)}{\sum_{t \in U_i} n(t, U_i)}. \quad (5)$$

Then an asymmetric topical similarity $\mathrm{TopicSim}(U_i, U_j)$ for utterances U_i to U_j (with direction $U_i \rightarrow U_j$) can be defined by accumulating $\mathrm{LTS}_t(T_k)$ in (2) weighted by $P(T_k \mid U_i)$ for all terms t in U_j over all latent topics:

$$\mathrm{TopicSim}(U_i, U_j) = \sum_{t \in U_j} \sum_{k=1}^{K} \mathrm{LTS}_t(T_k) P(T_k \mid U_i), \quad (6)$$

where the idea is very similar to the generative probability in IR. We call it generative significance of U_i given U_j.

Within a document d, the lexical similarity is the measure of word overlap between the utterance U_i and U_j. We compute $\mathrm{LexSim}(U_i, U_j)$ as the cosine similarity between two TF-IDF vectors from U_i and U_j like well-known LexRank (Erkan and Radev, 2004). Note that $\mathrm{LexSim}(U_i, U_j) = \mathrm{LexSim}(U_j, U_i)$

2.4 Intra-Speaker Topic Modeling

We assume a single speaker usually focuses on similar topics, so if an utterance is important, the scores of the utterances from the same speaker should be increased. Then we increase the similarity between the utterances from the same speaker to share the topics:

$$\mathrm{TopicSim}'_k(U_i, U_j) = \begin{cases} \mathrm{TopicSim}(U_i, U_j)^{1+w} \\ \text{, if } U_i \in S_k \text{ and } U_j \in S_k \\ \mathrm{TopicSim}(U_i, U_j)^{1-w} \\ \text{, otherwise} \end{cases} \quad (7)$$

where S_k is the set including all utterances from speaker k, and w is a weighting parameter for modeling the speaker relation, which means the level of coherence of topics within a single speaker. Here the topics from the same speaker can partially shared.

2.5 Integrated Random Walk

We modify random walk (Hsu and Kennedy, 2007; Chen et al., 2011) to integrate two types of similarity over the graph obtained above. $v(i)$ is the new score for node U_i, which is the interpolation of three scores, the normalized initial importance $r(i)$ for node U_i and the score contributed by all neighboring nodes U_j of node U_i weighted by $p_t(j,i)$ and $p_l(j,i)$,

$$\begin{aligned} v(i) &= (1-\alpha-\beta)r(i) \\ &+ \alpha \sum_{U_j \in B_i^t} p_t(j,i)v(j) + \beta \sum_{U_j \in B_i^l} p_l(j,i)v(j), \end{aligned} \tag{8}$$

where α and β are the interpolation weights, B_i^t is the set of neighbors connected to node U_i via topical incoming edges, B_i^l is the set of neighbors connected to node U_i via lexical incoming edges, and

$$r(i) = \frac{I(U_i,d)}{\sum_{U_j} I(U_j,d)} \tag{9}$$

is normalized importance scores of utterance U_i, $I(U_i,d)$ in (1). We normalize topical similarity by the total similarity summed over the set of outgoing edges, to produce the weight $p_t(j,i)$ for the edge from U_j to U_i on the graph. Similarly, $p_l(j,i)$ is normalized in lexical edges.

(8) can be iteratively solved with the approach very similar to that for the PageRank problem (Page et al., 1998). Let $\mathbf{v} = [v(i), i = 1,2,...,L]^\mathbf{T}$ and $\mathbf{r} = [r(i), i = 1,2,...,L]^\mathbf{T}$ be the column vectors for $v(i)$ and $r(i)$ for all utterances in the document, where L is the total number of utterances in the document d and $\mathbf{T}$ represents transpose. (8) then has a vector form below,

$$\begin{aligned} \mathbf{v} &= (1-\alpha-\beta)\mathbf{r} + \alpha\mathbf{P_t}\mathbf{v} + \beta\mathbf{P_l}\mathbf{v} \\ &= \left((1-\alpha-\beta)\mathbf{r}\mathbf{e}^\mathbf{T} + \alpha\mathbf{P_t} + \beta\mathbf{P_l}\right)\mathbf{v} = \mathbf{P'}\mathbf{v}, \end{aligned} \tag{10}$$

where $\mathbf{P_t}$ and $\mathbf{P_l}$ are $L \times L$ matrices of $p_t(j,i)$ and $p_l(j,i)$ respectively, and $\mathbf{e} = [1,1,...,1]^\mathbf{T}$. It has been shown that the solution $\mathbf{v}$ of (10) is the dominant eigenvector of $\mathbf{P'}$ (Langville and Meyer, 2006), or the eigenvector corresponding to the largest absolute eigenvalue of $\mathbf{P'}$. The solution $v(i)$ can then be obtained.

3 Experiments

3.1 Corpus

The corpus used in this research consists of a sequence of naturally occuring meetings, which featured largely overlapping participant sets and topics of discussion. For each meeting, SmartNotes (Banerjee and Rudnicky, 2008) was used to record both the audio from each participant as well as his notes. The meetings were transcribed both manually and using a speech recognizer; the word error rate is around 44%. In this paper we use 10 meetings held from April to June of 2006. On average each meeting had about 28 minutes of speech. Across these 10 meetings there were 6 unique participants; each meeting featured between 2 and 4 of these participants (average: 3.7). The total number of utterances is 9837 across 10 meetings. In this paper, we separate dev set (2 meetings) and test set (8 meetings). Dev set is used to tune the parameters such as α, β, w.

The reference summaries are given by the set of "noteworthy utterances": two annotators manually labelled the degree (three levels) of "noteworthiness" for each utterance, and we extract the utterances with the top level of "noteworthiness" to form the summary of each meeting. In the following experiments, for each meeting, we extract the top 30% number of terms as the summary.

3.2 Evaluation Metrics

Automated evaluation utilizes the standard DUC evaluation metric ROUGE (Lin, 2004) which represents recall over various n-grams statistics from a system-generated summary against a set of human generated peer summaries. F-measures for ROUGE-1 (unigram) and ROUGE-L (longest common subsequence) can be evaluated in exactly the same way, which are used in the following results.

3.3 Results

Table 1 shows the performance achieved by all proposed approaches. In these experiments, the damping factor, $(1-\alpha-\beta)$ in (8), is empirically set to 0.1. Row (a) is the baseline, which use LTE-based statistical measure to compute the importance of utterances $I(U,d)$. Row (b) is the result only considering lexical similarity; row (c) only uses topical similarity. Row (d) are the results additionally including speaker information such as $\mathrm{TopicSim}'(U_i,U_j)$. Row (e) is the result performed by integrated random walk (with $\alpha \neq 0$ and $\beta \neq 0$) using parameters that have been optimized on the dev set.

3.3.1 Graph-Based Approach

We can see the performance after graph-based recomputation, shown in rows (b) and (c), is significantly better than the baseline, shown in row (a), for both ASR and manual transcripts. For ASR transcripts, topical similarity and lexical similarity give similar results. For manual transcripts, topical similarity performs slightly worse than lexical similarity, because manual transcripts don't contain the recognition errors, and therefore word overlap can accurately measure the similarity between two utter-

F-measure		ASR Transcripts		Manual Transcripts	
		ROUGE-1	ROUGE-L	ROUGE-1	ROUGE-L
(a)	Baseline: LTE	46.816	46.256	44.987	44.162
(b)	LexSim ($\alpha = 0, \beta = 0.9$)	48.940	48.504	46.540	45.858
(c)	TopicSim ($\alpha = 0.9, \beta = 0$)	49.058	48.436	46.199	45.392
(d)	Intra-Speaker TopicSim	49.212	48.351	**47.104**	**46.299**
(e)	Integrated Random Walk	**49.792**	**49.156**	46.714	46.064
MAX RI		+6.357	+6.269	+4.706	+4.839

Table 1: Maximum relative improvement (RI) with respect to the baseline for all proposed approaches (%).

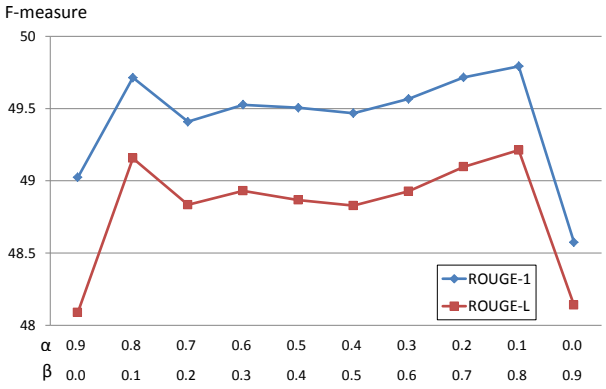

Figure 2: The performance from integrated random walk with different combination weights, α and β ($\alpha + \beta = 0.9$ in all cases) for ASR transcripts.

ances. However, for ASR transcripts, although topical similarity is not as accurate as lexical similarity, it can compensate for recognition errors, so that the approaches have similar performance. Thus, graph-based approaches can significantly improve the baseline results.

3.3.2 Effectiveness of Intra-Speaker Modeling

We find that modeling intra-speaker topics can improve the performance (row (c) and row (d)), which means speaker information is useful to model the topical similarity. The experiment shows intra-speaker modeling can help us include the important utterances for both ASR and manual transcripts.

3.3.3 Integration of Topical and Lexical Similarity

Row (e) shows the result of the proposed approach, which integrates topical and lexical similarity into a single graph, considering two types of relations together. For ASR transcripts, row (e) is better than row (b) and row (d), which means topical similarity and lexical similarity can model different types of relations, because of recognition errors. Figure 2 shows the sensitivity of the combination weights for integrated random walk. We can see topical similarity and lexical similarity are additive, i.e. they can compensate each other, improving the performance by integrating two types of edges in a single graph. Note that the exact values of α and β do not matter so much for the performance.

For manual transcripts, row (e) cannot perform better by combing two types of similarity, which means topical similarity can dominate lexical similarity, since without recognition errors topical similarity can model the relations accurately and additionally modeling intra-speaker topics can effectively improve the performance.

In addition, Banerjee and Rudnicky (2008) used supervised learning to detect noteworthy utterances on the same corpus, and achieved ROGURE-1 scores of around 43% for ASR, and 47% for manual transcriptions. Our unsupervised approach performs better, especially for ASR transcripts.

Note that the performance on ASR is better than on manual transcripts. Because a higher percentage of recognition errors occurs on "unimportant" words, these words tend to receive lower scores; we can then exclude the utterances with more errors, and achieve better summarization results. Other recent work has also demonstrated better performance for ASR than manual transcripts (Chen et al., 2011; Kong and Lee, 2011).

4 Conclusion and Future Work

Extensive experiments and evaluation with ROUGE metrics showed that intra-speaker topics can be modeled in topical similarity and that integrated random walk can combine the advantages from two types of edges for imperfect ASR transcripts, where we achieved more than 6% relative improvement. We plan to model inter-speaker topics in the graph-based approach in the future.

Acknowledgements

The first author was supported by the Institute of Education Science, U.S. Department of Education, through Grants R305A080628 to Carnegie Mellon University. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views or official policies, either expressed or implied of the Institute or the U.S. Department of Education.

References

Banerjee, S. and Rudnicky, A. I. 2008. An extractive-summarizaion baseline for the automatic detection of noteworthy utterances in multi-party human-human dialog. *Proc. of SLT*.

Chen, Y.-N., Huang, Y., Yeh, C.-F., and Lee, L.-S. 2011. Spoken lecture summarization by random walk over a graph constructed with automatically extracted key terms. *Proc. of InterSpeech*.

Erkan, G. and D. R. Radev., D. R. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, Vol. 22.

Furui, S., Kikuchi, T., Shinnaka, Y., and Hori, C. 2004. Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Trans. on Speech and Audio Processing*.

Garg, N., Favre, B., Reidhammer, K., and Hakkani-Tür 2009. ClusterRank: A graph based method for meeting summarization. *Proc. of InterSpeech*.

Gillick, D. J. 2011. The elements of automatic summarization. *PhD thesis, EECS, UC Berkeley*.

Glass J., Hazen, T. J., Cyphers, S., Malioutov, I., Huynh, D., and Barzilay, R. 2007. Recent progress in the MIT spoken lecture processing project. *Proc. of InterSpeech*.

Hofmann, T. 1999. Probabilistic latent semantic indexing. *Proc. of SIGIR*.

Hsu, W. and Kennedy, L. 2007. Video search reranking through random walk over document-level context graph. *Proc. of MM*.

Kong, S.-Y. and Lee, L.-S. 2011. Semantic analysis and organization of spoken documents based on parameters derived from latent topics. *IEEE Trans. on Audio, Speech and Language Processing*, 19(7): 1875-1889.

Langville, A. and Meyer, C. 2005. A survey of eigenvector methods for web information retrieval. *SIAM Review*.

Lee, L.-S. and Chen, B. 2005. Spoken document understanding and organization. *IEEE Signal Processing Magazine*.

Lin, C. 2004. Rouge: A package for automatic evaluation of summaries. *Proc. of Workshop on Text Summarization Branches Out*.

Liu, F. and Liu, Y. 2010. Using spoken utterance compression for meeting summarization: A pilot study. *Proc. of SLT*.

Liu Y., Xie, S., and Liu, F. 2010. Using N-best recognition output for extractive summarization and keyword extraction in meeting speech. *Proc. of ICASSP*.

Page, L., Brin, S., Motwani, R., Winograd, T. 1998. The pagerank citation ranking: bringing order to the web. *Technical Report, Stanford Digital Library Technologies Project*.