Integrating Intra-Speaker Topic Modeling and Temporal-Based Inter-Speaker Topic Modeling in Random Walk for Improved Multi-Party Meeting Summarization

Yun-Nung Chen and Florian Metze

School of Computer Science, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213-3891, USA
{yvchen, fmetze}@cs.cmu.edu

Abstract

This paper proposes an improved approach of summarization for spoken multi-party interaction, in which intra-speaker and inter-speaker topics are modeled in a graph constructed with topical relations. Each utterance is represented as a node of the graph and the edge between two nodes is weighted by the similarity between the two utterances, which is topical similarity evaluated by probabilistic latent semantic analysis (PLSA). We model intra-speaker topics by sharing the topics from the same speaker and inter-speaker topics by partially sharing the topics from the adjacent utterances based on temporal information. We did experiments for ASR and manual transcripts. For both types of transcripts, experiments confirmed the efficacy of combining intra- and inter-speaker topic modeling for summarization.

Index Terms: summarization, PLSA, topic transition, temporal information

1. Introduction

Speech summarization is very important [1], because multimedia/spoken documents are more difficult to browse, and it has been actively investigated before. While most work focused primarily on news content, recent effort has been increasingly directed to new domains such as lectures [2, 3] and multi-party interaction [4, 5, 6]. We take meeting recording as multi-party interaction and do experiments on this dataset, where we perform extractive summarization on ASR and manual transcripts [7].

For text summarization, many approaches focus on graph-based methods to compute lexical centrality of each utterance to extract summaries [8]. The 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 [9], 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:

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

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, we proposed a graphical structure to rescore $I(U, d)$ above in (1), which can model the topical coherence between utterances using random walk within documents [3, 5]. Unlike lecture and news summarization, meeting recording is the multi-party interaction corpus so that the relations such as topic distribution within a single speaker or between speakers can be considered. Thus, this paper models intra- and inter-speaker topics together in the graph by partially sharing topics with the utterances from the same speaker or adjacent utterances to improve meeting summarization [10].

2. Proposed Approach

We first preprocess the utterances in all meetings: word stemming and noise utterance filtering. Then we construct a graph to compute the importance of all utterances. 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 similarity. The basic idea is that an utterance similar to more important utterances should be more important [3]. 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 is in Figure 1, in which A_i and B_i are the sets of neighbors of the node U_i connected respectively by outgoing and incoming edges.

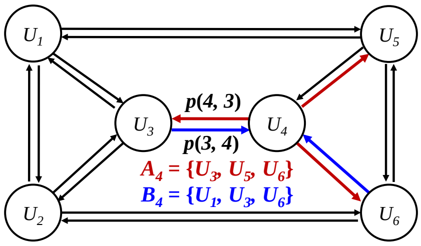

Figure 1: A simplified example of the graph considered.

2.1. Parameters from Topic Model

Probabilistic latent semantic analysis (PLSA) [11] 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 [11]. We utilize two parameters from PLSA, latent topic significance (LTS) and latent topic entropy (LTE) [12]. The parameters also can be computed by other topic model such as latent dirichlet allocation(LDA) [13] in similar way.

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

Here in this work, the statistical measure of a term t_i, $s(t_i, d)$ in (1) can be defined based on $\mathrm{LTE}(t_i)$ 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 $0 \leq s(t_i, d) \leq 1$, so the score $s(t_i, d)$ is inversely proportion to the latent topic entropy $\mathrm{LTE}(t_i)$. Some works [12] showed that this measure outperformed the very successful "significance score" [9] in speech summarization, and here we use LTE-based statistical measure, $s(t_i, d)$, as the baseline.

2.3. Topical 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 $\mathrm{Sim}(U_i, U_j)$ for utterances U_i to U_j (with direction $U_i \to U_j$) 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,

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

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

2.4. Intra/Inter-Speaker Topic Modeling

We additionally consider speaker information to model topics more accurately,

$$\mathrm{Sim}'(U_i, U_j) = \mathrm{Sim}(U_i, U_j)^w, \quad (7)$$

$$w = 1 + w_{intra}(U_i, U_j) + w_{inter}(U_i, U_j), \quad (8)$$

where w_{intra} is topic sharing weight for intra-speaker and w_{inter} is for inter-speaker topic sharing, which are described as Section 2.4.1 and 2.4.2 respectively.

2.4.1. Intra-Speaker Topic Sharing Weight

Since we assume that the utterances from the same speaker in the dialogue usually focus on similar topics, which means if an utterance is important, the other utterances from the same speaker are more likely to be important in the dialogue [5]. Then we can estimate $\mathrm{Sim}'(U_i, U_j)$ by setting $w_{intra}(U_i, U_j)$ as

$$w_{intra}(U_i, U_j) = \begin{cases} +\delta & \text{, if } U_i \in S_k \text{ and } U_j \in S_k \\ -\delta & \text{, otherwise} \end{cases} \quad (9)$$

S_k is the set including all utterances from speaker k and δ is a weighting parameter for modeling the speaker relation. Here the topics from the same speaker can partially shared.

2.4.2. Inter-Speaker Topic Sharing Weight

Topic transition between adjacent utterances should be slow so that adjacent utterances should have similar topic distribution [14] even though they are not from the same speaker, and then we can increase $\mathrm{Sim}'(U_i, U_j)$ if U_i and U_j have closer position in the dialogue. Thus, we compute the weight for inter-speaker topic sharing as

$$w_{inter}(U_i, U_j) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(l_j - l_i)^2}{2\sigma^2}), \quad (10)$$

where l_i is the position of the utterance U_i in the dialogue, which means U_i is the l_i-th utterance in the dialogue. The boundary of utterance is decided by SmartNote [4]. (10) is under an assumption that topic sharing is based on a normal distribution with a standard deviation σ. If $|l_i - l_j|$ is smaller, which means U_i and U_j is closer to each other, and they may share their topics so that $w_{inter}(U_i, U_j)$ is larger in (10). σ is a parameter of topic sharing range, which can be tuned by dev set.

We normalize the similarity summed over the top N utterance U_k with edges outgoing from U_i, or the set A_i, to produce the weight $p(i, j)$ for the edge from U_i to U_j on the graph,

$$p(i,j) = \frac{\mathrm{Sim}'(U_i, U_j)}{\sum_{U_k \in A_i} \mathrm{Sim}'(U_i, U_k)}. \qquad (11)$$

2.5. Random Walk

We use random walk [3, 15] to integrate two types of scores over the graph obtained above. $v(i)$ is the new score for node U_i, which is the interpolation of two 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 $p(j, i)$,

$$v(i) = (1-\alpha) r(i) + \alpha \sum_{U_j \in B_i} p(j,i) v(j), \qquad (12)$$

where α is the interpolation weight, B_i is the set of neighbors connected to node U_i via incoming edges, and $r(i)$ is normalized importance scores of utterance U_i, $I(U_i, d)$ in (1).

(12) can be iteratively solved with the approach very similar to that for the PageRank problem [16]. 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. (12) then has a vector form below,

$$\begin{aligned} \mathbf{v} &= (1-\alpha)\mathbf{r} + \alpha \mathbf{P} \mathbf{v} \qquad (13) \\ &= \left((1-\alpha)\mathbf{r}\mathbf{e}^{\mathbf{T}} + \alpha \mathbf{P}\right) \mathbf{v} = \mathbf{P}'\mathbf{v}, \end{aligned}$$

where $\mathbf{P}$ is $L \times L$ matrices of $p(j, i)$, and $\mathbf{e} = [1, 1, ..., 1]^{\mathbf{T}}$. Because $\sum_i v(i) = 1$ from (12), $\mathbf{e}^{\mathbf{T}}\mathbf{v} = 1$. It has been shown that the closed-form solution $\mathbf{v}$ of (13) is the dominant eigenvector of $\mathbf{P}'$ [17], 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 is the sequences of natural meetings, which featured largely overlapping participant sets and topics of discussion. For each meeting, SmartNotes [4] 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). 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, σ, and δ.

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 following experiments, for each meeting, we extract top 30% number of terms as the summary.

3.2. Evaluation Metrics

Automated evaluation will utilize the standard DUC evaluation metric ROUGE [18] 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 from all proposed approaches. 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 after applying random walk with only topical similarity. Row (c) is the result additionally including intra-speaker topic modeling ($w_{intra} \neq 0$); row (d) includes inter-speaker topic modeling ($w_{inter} \neq 0$). Row (e) is the result performed by integrating two types of speaker information (with $w_{intra} \neq 0$ and $w_{inter} \neq 0$).

Note that the performance of ASR is better than manual transcripts. Because a higher percentage of errors is on "unimportant" words, the recognition errors are harder to obtain high scores; then we can exclude the utterances with more errors to get better summarization results. Some recent works also show better performance for ASR than manual transcripts [3, 5].

3.3.1. Graph-Based Approach

We can see the performance after graph-based recomputation row (b) is significantly better than baseline row (a) for both ASR and manual transcripts. The improvement for ASR is more than for manual transcripts,

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)	Random Walk	49.058	48.436	46.199	45.392
(c)	Random Walk + Intra-Speaker	49.212	48.351	47.104	46.299
(d)	Random Walk + Inter-Speaker	48.927	48.305	46.291	45.481
(e)	Random Walk + Inter-Speaker + Intra-Speaker	**49.640**	**48.865**	**48.091**	**47.364**
MAX RI		+6.032	+5.640	+6.900	+7.251

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

because ASR contains some recognition errors, which makes original scores measured inaccurately, and random walk is used to propagate importance based on topical similarity can compensate recognition errors. Thus, graph-based approaches can significantly improve the baseline results.

3.3.2. Effectiveness of Speaker Information Modeling

We find that modeling intra-speaker topics can improve the performance (row (b) and row (c)), 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. Then we find that only modeling inter-speaker topics cannot offer significant improvement for ASR transcripts (row (b) and row (d)) probably because sharing topics with adjacent utterances may decrease the centrality especially for the utterances with recognition errors. For manual transcripts, the improvement of inter-speaker topic model is not significant.

Row (e) is the result from proposed approach, which integrates intra-speaker and inter-speaker topic modeling into a single graph, considering two types of relations together. For ASR transcripts, row (e) is better than row (c) and row (d), which means intra-speaker and inter-speaker information cover different types of relations, and the relations can be additive. Note that only using inter-speaker topic modeling cannot improve the performance, but integrating with intra-speaker topic modeling can offer better results. The reason may be that intra-speaker topic modeling enhances centrality of important utterances, and additionally involving inter-speaker topic modeling slightly decreases centrality but successfully smoothing topic transition for adjacent utterances. For manual transcripts, row (e) also perform better by combing two types of speaker information, and the improvement is larger than ASR transcripts. Since without recognition errors topical similarity can model the relations accurately, integrating two types of speaker information can effectively improve the performance.

In addition, Banerjee and Rudnicky [4] used supervised learning to detect noteworthy utterances in the same corpus, performing 43% (ASR) and 47% (manual) for ROUGE-1. Compared to it, our unsupervised approach performs better especially for ASR transcripts.

4. Conclusions

Extensive experiments and evaluation with ROUGE metrics showed that inter- and intra-speaker topics can be modeled together in one single graph and that random walk can combine the advantages from two types of speaker information for both ASR and manual transcripts, where we achieved more than 6% relative improvement.

5. References

[1] L. Lee and B. Chen. “Spoken document understanding and organization”, in *IEEE Signal Processing Magazine*, 2005.

[2] J. Glass et al., “Recent progress in the MIT spoken lecture processing project”, in *Proc. of InterSpeech*, 2007.

[3] Y. Chen et al., “Spoken lecture summarization by random walk over a graph constructed with automatically extracted key terms”, in *Proc. of InterSpeech*, 2011.

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

[5] Y. Chen and F. Metze, “Intra-speaker topic modeling for improved multi-party meeting summarization with integrated random walk”, in *Proc. of NAACL-HLT*, 2012.

[6] F. Liu and Y. Liu, “Using spoken utterance compression for meeting summarization: A pilot study”, in *Proc. of SLT*, 2010.

[7] Y. Liu et al., “Using N-best recognition output for extractive summarization and keyword extraction in meeting speech”, in *Proc. of ICASSP*, 2010.

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

[9] S. Furui et al., “Speech-to-text and speech-to-speech summarization of spontaneous speech”, in *IEEE Trans. on Speech and Audio Processing*, 2004.

[10] N. Garg et al., “ClusterRank: A graph based method for meeting summarization”, in *Proc. of InterSpeech*, 2009.

[11] T. Hofmann, ”Probabilistic latent semantic indexing”, in *Proc. of SIGIR*, 1999.

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

[13] D. M. Blei et al., “Latent dirichilet allocation”, in *Journal of Machine Learning Research*, 2003.

[14] H. Lee et al., “Utterance-level latent topic transition modeling for spoken documents and its application in automatic summarization”, in *ICASSP*, 2012.

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

[16] L. Page et al., “The pagerank citation ranking: bringing order to the web”, in *Technical Report, Stanford Digital Library Technologies Project*, 1998.

[17] A. Langville and C. Meyer, “A survey of eigenvector methods for web information retrieval”, in *SIAM Review*, 2005.

[18] C. Lin, “Rouge: A package for automatic evaluation of summaries”, in *Proc. of Workshop on Text Summarization Branches Out*, 2004.