ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

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

Language Technologies Institute, 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 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 the topical similarity, as 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. For both manual transcripts and ASR output, experiments confirmed the efficacy of combining intra- and inter-speaker topic modeling for summarization.

Index Terms: summarization, multi-party meeting, topic model, probabilistic latent semantic analysis (PLSA), topic transition, temporal information, random walk

1. Introduction

Speech summarization is important [1] for spoken or even multimedia documents, which are more difficult to browse than text, and has therefore been investigated in the past. While most work focused primarily on news content, recent effort has been increasingly directed towards new domains such as lectures [2, 3] and multi-party interaction [5, 6, 7]. In this work, we perform extractive summarization on the output of automatic speech recognition (ASR) and corresponding manual transcripts [8] of multi-party "meeting" recordings.

Many approaches to text summarization focus on graph-based methods to compute lexical centrality of each utterance, in order to extract summaries [9]. Speech summarization carries intrinsic difficulties due to the presence of recognition errors, spontaneous speech effects, and lack of segmentation. A general approach has been found to be very successful [10], 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

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

where $s(t_i,d)$, $l(t_i)$, $c(t_i)$, and $g(t_i)$ respectively are some statistical measure (such as TF-IDF), some linguistic measure (e.g., different part-of-speech tags are given different weights), a confidence score, and an N-gram score for the term t_i; $b(U)$ 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.

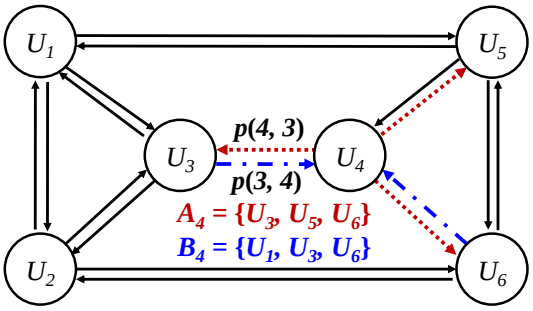

Figure 1: *A simplified example of the graph considered.*

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 a random walk process within documents [3, 6]. Unlike lecture and news summarization, meeting recordings contain spoken multi-party interactions, 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 [11].

2. Proposed Approach

We first preprocess the utterances in all meetings by applying word stemming[1] and noise utterance filtering. Then we construct a graph to compute the importance of all utterances. We formulate the utterance selection problem as a random walk on a directed graph, in which each utterance is a node and the edges between these are weighted by topical similarity. The basic idea is that an utterance similar to more important utterances should be more important [3, 4]. We then keep only the top N outgoing edges with the highest weights from each node, while considering incoming edges to each node for importance propagation in the graph. Figure 1 shows a simplified example for such a

[1] http://www.tartarus.org/ martin/PorterStemmer

graph, in which A_i and B_i are the sets of neighbors of the node U_i, connected by outgoing and incoming edges respectively.

2.1. Parameters from Topic Model

Probabilistic latent semantic analysis (PLSA) [12] 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 using the EM algorithm, by maximizing a likelihood function [12]. We utilize two parameters from PLSA, latent topic significance (LTS) and latent topic entropy (LTE) [13]. The parameters can also be computed by other topic models, such as latent dirichilet allocation (LDA) [14] in a 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 that 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

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 $s(t_i, d)$ lies within the interval $[0, 1]$, so the score $s(t_i, d)$ is inversely proportion to the latent topic entropy $\mathrm{LTE}(t_i)$. In [13], this measure outperformed the very successful "significance score" [10] in speech summarization, so we use the LTE-based statistical measure, $s(t_i, d)$, as our 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 $U_i \rightarrow U_j$) 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,

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

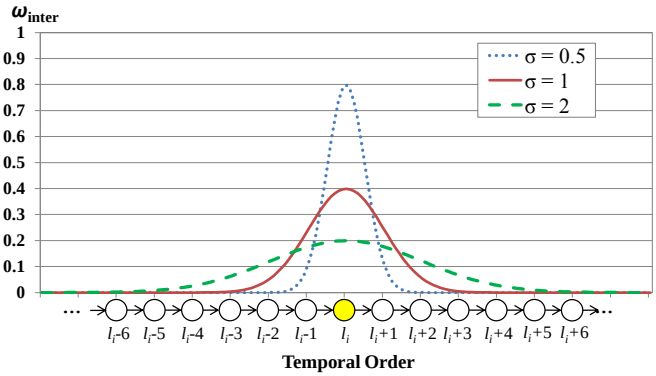

Figure 2: *The inter-speaker topic sharing weights, w_{inter}, related to the l_i-th utterance in terms of different topic sharing range parameters σ. Smaller σ means utterances share their topic distribution to less temporally adjacent utterances.*

where the idea is similar to generative probability in information retrieval. 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,

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

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

where w is the weight for modeling intra- and inter-speaker topics. w_{intra} is the intra-speaker topic sharing weight and w_{inter} is the inter-speaker topic sharing weight, as described below.

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, this means that if an utterance is important, the other utterances from the same speaker are more likely to be important in the dialogue [6] as well. We can then estimate $\mathrm{Sim}'(U_i, U_j)$ by setting $w_{\text{intra}}(U_i, U_j)$ as

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

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

2.4.2. Temporal-Based Inter-Speaker Topic Sharing Weight

Topic transition between temporally adjacent utterances should be slow, so that temporally adjacent utterances should have similar topic distribution [15], even though they are not from the same speaker. We can then increase $\mathrm{Sim}'(U_i, U_j)$ if U_i and U_j have a closer position in the dialogue. Thus, we compute the weight for inter-speaker topic sharing as

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

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. In our implementation, the boundary of an utterance is set by SmartNote [5]. (10) assumes that topic sharing is based on a normal distribution with a standard deviation σ.

Figure 2 shows the topic sharing weights related to the l_i-th utterance based on normal distribution with different σ. If $|l_i - l_j|$ is smaller, which means U_i and U_j are closer to each other, they may share their topics so that $w_{\text{inter}}(U_i, U_j)$ is larger in (10). σ is a topic sharing range parameter, which can be tuned on the development 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 in the graph,

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

2.5. Random Walk

We use random walk [3, 16] to integrate the 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 $p(j, i)$,

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

where α is the interpolation weight, B_i is the set of neighbors connected to node U_i via incoming edges, and

$$r(i) = \frac{I(U_i, d)}{\sum_{U_j} I(U_j, d)} \quad (13)$$

is the normalized importance scores of utterance U_i, $I(U_i, d)$ in (1).

(12) can be iteratively solved with an approach very similar to the PageRank algorithm [17]. 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 a transposition. (12) then has a vector form below,

$$\begin{aligned} \mathbf{v} &= (1-\alpha)\mathbf{r} + \alpha\mathbf{P}\mathbf{v} \quad (14) \\ &= \left((1-\alpha)\mathbf{r}\mathbf{e}^{\mathbf{T}} + \alpha\mathbf{P}\right)\mathbf{v} = \mathbf{P}'\mathbf{v}, \end{aligned}$$

where the $\mathbf{P}$ are $L \times L$ 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 (14) is the dominant eigenvector of $\mathbf{P}'$ [18], 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 a sequences of natural meetings, which features largely overlapping participant sets and topics of discussion. For each meeting, SmartNotes [5] 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 use a separate development set (2 meetings) and test set (8 meetings). The development 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 highest level of "noteworthiness" to form the summary of each meeting. In the following experiments, for each meeting, we extract about 30% of the number of terms as the summary.

3.2. Evaluation Metrics

Our automated evaluation utilizes the standard DUC evaluation metric, ROUGE [19], which represents recall over various n-grams statistics from a system-generated summary against a set of human generated summaries. F-measures for ROUGE-1 (unigram) and ROUGE-L (longest common subsequence) can be evaluated in exactly the same way.

3.3. Results

Table 1 shows the performance achieved from all proposed approaches. Row (a) is the baseline, which uses an 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 ($w_{\text{intra}} \neq 0$). Row (d) includes inter-speaker topic modeling ($w_{\text{inter}} \neq 0$). Row (e) is the result performed by integrating two types of speaker information (with $w_{\text{intra}} \neq 0$ and $w_{\text{inter}} \neq 0$).

Note that the performance of ASR is better than manual transcripts. Because a higher percentage of errors is on "unimportant" words, incorrectly recognized words find it harder to obtain high scores; so utterances with more errors tend to get excluded from the summarization results. Other recent work also shows better performance for ASR than manual transcripts [3, 6].

3.3.1. Graph-Based Approach

We can see the performance after graph-based re-computation (row (b)) is significantly better than baseline (row (a)) for both ASR and manual transcripts. The improvement for ASR is larger than for manual transcripts, because ASR output contains recognition errors, which makes determination of original scores inaccurate, and random walk is used to propagate importance based on topical similarity, which can effectively compensate recognition errors. Thus, graph-based approaches can significantly improve on the baseline results.

3.3.2. Intra-Speaker Information Modeling

We find that modeling intra-speaker topics improves performance (see row (b) and row (c)), which means the utterances from the speakers who speak more important utterances tend to be more important. Thus, propagating the importance scores between the utterances from the same speaker can improve the results. The experiment shows intra-speaker modeling can help include the important utterances for both ASR and manual transcripts.

3.3.3. Inter-Speaker Topic Modeling

We also find that only modeling inter-speaker topics cannot offer significant improvement for ASR transcripts (row (b) and

Table 1: *The results of all proposed approaches and maximum relative improvement with respect to the baseline (%).*

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 Relative Improvement		+6.032	+5.640	+6.900	+7.251

row (d)), probably because sharing topics with temporally 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.

3.3.4. Integration Intra- and Inter-Speaker Topic Modeling

Row (e) shows the result from the 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- 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 integration 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 smoothes topic transitions between temporally adjacent utterances. For manual transcripts, row (e) also performs better by combing two types of speaker information, and the improvement is larger than for ASR transcripts. Since in the absence of recognition errors topical similarity can model the relations accurately, integrating two types of speaker information can effectively improve the performance.

On the same corpus, Banerjee and Rudnicky [5] used supervised learning to detect noteworthy utterances in the same corpus, achievieng ROUGE-1 scores of 43% (ASR) and 47% (manual). In comparison, our unsupervised approach performs better, especially for ASR transcripts.

4. Conclusions and Future Work

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. In the future, we plan to modify the graph into a two-level graph to model speakers' topics and utterances' topics in different levels.

5. 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. We thank the students, educators who helped create our data, and the reviewers for their helpful comments.

6. References

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

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

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

[4] Chen, Y.-N. et al., "Improved spoken term detection with graph-based re-ranking in feature space", in *IEEE ICASSP*, 2011.

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

[6] Chen, Y.-N. and Metze, F., "Intra-speaker topic modeling for improved multi-party meeting summarization with integrated random walk", in *NAACL-HLT*, 2012.

[7] Liu, F. and Liu, Y., "Using spoken utterance compression for meeting summarization: A pilot study", in *IEEE SLT*, 2010.

[8] Liu, Y. et al., "Using N-best recognition output for extractive summarization and keyword extraction in meeting speech", in *IEEE ICASSP*, 2010.

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

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

[11] Garg, N. et al., "ClusterRank: A graph based method for meeting summarization", in *InterSpeech*, 2009.

[12] Hofmann, T., "Probabilistic latent semantic indexing", in *SIGIR*, 1999.

[13] Kong, S.-Y. and Lee, L.-S., "Semantic analysis and organization of spoken documents based on parameters derived from latent topics", in *IEEE Trans. on Audio, Speech and Language Processing*, 2011.

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

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

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

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

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

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