TWO-LAYER MUTUALLY REINFORCED 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 a two-layer graph with utterance-to-utterance, speaker-to-speaker, and speaker-to-utterance relations is constructed. Each utterance and each speaker are represented as a node in the utterance-layer and speaker-layer of the graph respectively, and the edge between two nodes is weighted by the similarity between the two utterances, the two speakers, or the utterance and the speaker. The relation between utterances is evaluated by lexical similarity via word overlap or topical similarity via probabilistic latent semantic analysis (PLSA). By within- and between-layer propagation in the graph, the scores from different layers can be mutually reinforced so that utterances can automatically share the scores with the utterances from the same speaker and similar utterances. For both ASR output and manual transcripts, experiments confirmed the efficacy of involving speaker information in the two-layer graph for summarization.

***Index Terms*—** Summarization, multi-party meeting, mutual reinforcement, 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 [4, 5, 6]. In this work, we perform extractive summarization on the output of automatic speech recognition (ASR) and corresponding manual transcripts [7] 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 [8, 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

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

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.

In recent work, we proposed a graphical structure to rescore $I(U, d)$ in (1) above, which can model the topical coherence between utterances using a random walk process within documents [3, 5]. Unlike lecture and news summarization, meeting recordings contain spoken multi-party interactions, so that the speaker "importance" scores can be added to the estimation of the importance of individual utterance [11]. Thus, this paper proposes to use two-layer mutually reinforced random walk to compute the speaker importance and to increase the scores of utterances similar to the utterances from important speakers. It models intra- and inter-speaker topics together in the two-layer graph by automatically propagating scores to the utterances from the same speaker or similar utterances to improve meeting summarization [9, 12].

Section 2 describes the construction of two-layer graph and the algorithms about computing the importance of utterances, which includes between-layer propagation and integration of within- and between-layer propagation. Section 3 shows the results of applying proposed approaches, and discusses the difference between two algorithms and the difference between topical and lexical similarity for both ASR and manual transcripts. Section 4 concludes.

2. PROPOSED APPROACH

In this paper, we use ASR and manual transcripts and make both types of text similar. We first preprocess the utterances in all meetings by applying word stemming[1] and noise utterance filtering, where the utterances with word counts smaller than 3 are removed. For extractive summarization, we set a cut-off ratio to retain only the most important utterances to form the summary of each document based on the "importance" of each utterance. Thus, we formulate the utterance selection problem as computing the importance of each utterance. Then we construct a two-layer graph to compute the importance for all utterances and speakers in speaker-layer and utterance-layer respectively. In the two-layer directed graph, each utterance is a node in utterance-layer and the edges between these are weighted by topical or lexical similarity described in Section 2.3. Each speaker in the meeting is a node in speaker-layer and the edges between them are weighted by speaker-to-speaker relation. The edges between different layers are weighted by the relation between speakers and utterances.

The basic idea is that an utterance similar to more important utterances should be more important [3, 13], so the importance of each utterance considers the scores propagated from other utterances according to the similarity between them. In this approach, the propagated scores are not only based on utterance-to-utterance relation. Instead, the scores integrate three types of relations (utterance-to-utterance, speaker-to-speaker, and utterance-to-speaker) to automatically consider speaker information in the graph. Figure 1 shows a simplified example for such a two-layer graph, in which there are speaker-layer and utterance-layer containing speaker nodes and utterance nodes respectively.

2.1. Parameters from Topic Model

Probabilistic latent semantic analysis (PLSA) [14] 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 [14]. We utilize two parameters from PLSA, latent topic significance (LTS) and latent topic entropy (LTE) [15]. The parameters can also be computed by other topic models, such as latent dirichilet allocation (LDA) [16] 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

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

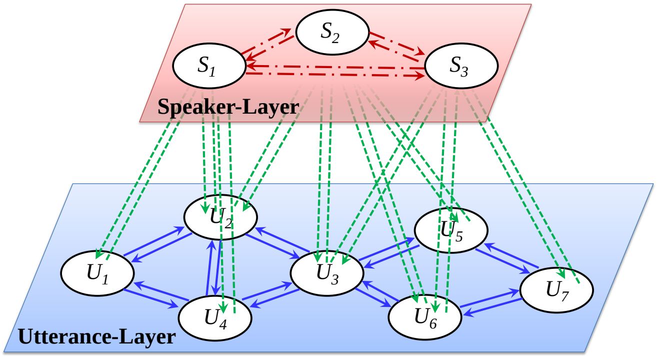

Fig. 1. *A simplified example of the two-layer graph considered, where a speaker S_i is represented as a node in speaker-layer and an utterance U_j is represented as a node in utterance-layer of the two-layer graph. There are three different types of edges corresponding to different relations (utterance-to-utterance, speaker-to-speaker, and utterance-to-speaker). Note that each utterance node has edges connected to all speaker nodes not only the sourcing speaker node.*

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 $P(T_k \mid t_i)$,

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

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) measures the importance of t_i such as TF-IDF. In this work, it can be defined based on $\mathrm{LTE}(t_i)$ in (3) as

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

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 $\mathrm{LTE}(t_i)$. This measure outperformed the very successful "significance score" [15, 10] in speech summarization, so we use the LTE-based statistical measure, $s(t_i, d)$, as our baseline.

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

2.3. Similarity between Utterances

We compute two different types of similarity between utterances based on topical and lexical distribution.

2.3.1. *Topical Similarity via PLSA*

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

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

Then an asymmetric topical similarity $\mathrm{TopicSim}(U_i, U_j)$ 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,

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

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

2.3.2. *Lexical Similarity via Word Overlap*

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 [8]. Note that $\mathrm{LexSim}(U_i, U_j) = \mathrm{LexSim}(U_j, U_i)$.

2.4. Two-Layer Mutually Reinforced Random Walk

For each document d, we construct a linked two-layer graph G containing utterance set and speaker set to compute the importance of each utterance. $G = \langle V_U, V_S, E_{UU}, E_{SS}, E_{US} \rangle$, where $V_U = \{U_i \in d\}$, $V_S = \{S_i \in d\}$, $E_{UU} = \{e_{ij} \mid U_i, U_j \in V_U\}$, $E_{SS} = \{e_{ij} \mid S_i, S_j \in V_S\}$, and $E_{US} = \{e_{ij} \mid U_i \in V_U, S_j \in V_S\}$. E_{UU}, E_{SS}, and E_{US} correspond the relation between utterances, the relation between speakers, and the relation between utterances and speakers respectively [9].

We compute $L_{UU} = [w_{U_i,U_j}]_{|V_U| \times |V_U|}$, where w_{U_i,U_j} is either from $\mathrm{TopicSim}(U_i, U_j)$ or $\mathrm{LexSim}(U_i, U_j)$. Word overlap between utterances may be sparse due to recognition errors, so it's possible that topical similarity via PLSA can capture more information than lexical similarity. $L_{SS} = [w_{S_i,S_j}]_{|V_S| \times |V_S|}$, where w_{S_i,S_j} is the cosine similarity between the TF-IDF vectors containing all utterances from speaker S_i and S_j, which means a speaker node in the graph is represented by all utterances from the speaker. Similarly, $L_{US} = [w_{U_i,S_j}]_{|V_U| \times |V_S|}$ and $L_{SU} = [w_{U_j,S_i}]_{|V_S| \times |V_U|}$, where w_{U_i,S_j} is the TF-IDF cosine similarity between utterance vector and speaker vector. Note that it is possible that $w_{U_i,S_j} > 0$ when $U_i \notin S_j$ because they both may have the same terms. Row-normalization are performed for L_{UU}, L_{SS}, L_{US}, L_{SU} [17]. They can be viewed as utterance-to-utterance, speaker-to-speaker, and utterance-to-speaker affinity metrics. Note that L_{US} is different from L_{SU}^T because of row-normalization.

Traditional random walk integrates the original scores and the scores propagated from other utterance nodes [3, 11, 18]. Here the proposed approach additionally considers the speaker information and integrates importance propagated from speaker nodes to model intra- and inter-speaker relation automatically. We propose two algorithms, one using scores only from between-layer propagation and another using scores both from within- and between-layer propagation, which are described in Section 2.4.1 and 2.4.2.

2.4.1. *Between-Layer Propagation*

Here we use two-layer mutually reinforced random walk to propagate the scores based on external mutual reinforcement between different layers through the edges E_{US}.

$$\begin{cases} F_U^{(t+1)} = (1-\alpha) F_U^{(0)} + \alpha \cdot L_{US} F_S^{(t)} \\ F_S^{(t+1)} = (1-\alpha) F_S^{(0)} + \alpha \cdot L_{SU} F_U^{(t)} \end{cases} \qquad (7)$$

$F_U^{(t)}$ and $F_S^{(t)}$ denote the importance scores of the utterance set V_U and speaker set V_S in t-th iteration respectively. In the algorithm, they are the interpolations of two scores, the initial importance ($F_U^{(0)}$ and $F_S^{(0)}$) and the scores propagated from another layer (speaker-layer and utterance-layer).

For utterance set, each utterance combines initial importance and the scores propagated from speaker-layer weighted by utterance-to-speaker similarity. Similarly, nodes of speaker-layer also include the scores propagated from utterance-layer. Then $F_U^{(t+1)}$ and $F_S^{(t+1)}$ can be mutually updated by the latter parts in (7) iteratively.

The algorithm will converge and then (8) can be satisfied [9].

$$\begin{cases} F_U^{*} = (1-\alpha) F_U^{(0)} + \alpha \cdot L_{US} F_S^{*} \\ F_S^{*} = (1-\alpha) F_S^{(0)} + \alpha \cdot L_{SU} F_U^{*} \end{cases} \qquad (8)$$

We can solve F_U^* as below.

$$\begin{aligned} F_U^* &= (1-\alpha) F_U^{(0)} \qquad (9) \\ &+ \alpha \cdot L_{US} \left((1-\alpha) F_S^{(0)} + \alpha \cdot L_{SU} F_U^* \right) \\ &= (1-\alpha) F_U^{(0)} + \alpha(1-\alpha) L_{US} F_S^{(0)} \\ &+ \alpha^2 L_{US} L_{SU} F_U^* \\ &= \Big((1-\alpha) F_U^{(0)} e^T + \alpha(1-\alpha) L_{US} F_S^{(0)} e^T \\ &+ \alpha^2 L_{US} L_{SU} \Big) F_U^* \\ &= M_1 F_U^*, \end{aligned}$$

where the $e = [1, 1, ..., 1]^T$. It has been shown that the closed-form solution F_U^* of (9) is the dominant eigenvector of M_1 [19], or the eigenvector corresponding to the largest absolute eigenvalue of M_1. The solution of F_U^* denotes the updated importance scores for all utterances. Similar to the PageRank algorithm [20], the solution can also be obtained by iteratively updating $F_U^{(t)}$ and $F_S^{(t)}$.

2.4.2. *Integrating Within- and Between-Layer Propagation*

Here we use a two-layer mutually reinforced random walk to propagate the scores based on internal importance propagation within the same layer and external mutual reinforcement between different layers.

$$\begin{cases} F_U^{(t+1)} = (1-\alpha)F_U^{(0)} + \alpha \cdot L_{UU}^T L_{US} F_S^{(t)} \\ F_S^{(t+1)} = (1-\alpha)F_S^{(0)} + \alpha \cdot L_{SS}^T L_{SU} F_U^{(t)} \end{cases} \tag{10}$$

Here $F_U^{(t)}$ and $F_S^{(t)}$ integrates the initial importance and the score including within- and between-layer propagation.

For utterance set, $L_{US}F_S^{(t)}$ is the score from speaker set weighted by utterance-to-speaker similarity, and then the scores are propagated based on utterance-to-utterance similarity L_{UU}. Compared to Section 2.4.1, the algorithm additionally considers the within-layer relation through L_{UU} and L_{SS}. Then $F_U^{(t+1)}$ can be updated by the latter part in (10), and $F_S^{(t+1)}$ as well. Similarly, the algorithm converges satisfying (11).

$$\begin{cases} F_U^* = (1-\alpha)F_U^{(0)} + \alpha \cdot L_{UU}^T L_{US} F_S^* \\ F_S^* = (1-\alpha)F_S^{(0)} + \alpha \cdot L_{SS}^T L_{SU} F_U^* \end{cases} \tag{11}$$

$$\begin{aligned} F_U^* &= (1-\alpha)F_U^{(0)} \\ &+ \alpha \cdot L_{UU}^T L_{US} \Big((1-\alpha)F_S^{(0)} + \alpha \cdot L_{SS}^T L_{SU} F_U^*\Big) \\ &= (1-\alpha)F_U^{(0)} + \alpha(1-\alpha)L_{UU}^T L_{US} F_S^{(0)} \\ &+ \alpha^2 L_{UU}^T L_{US} L_{SS}^T L_{SU} F_U^* \\ &= \Big((1-\alpha)F_U^{(0)} e^T + \alpha(1-\alpha)L_{UU}^T L_{US} F_S^{(0)} e^T \\ &+ \alpha^2 L_{UU}^T L_{US} L_{SS}^T L_{SU}\Big) F_U^* \\ &= M_2 F_U^*. \end{aligned} \tag{12}$$

Similarly, the closed-form solution F_U^* of (9) is the dominant eigenvector of M_2 [19].

For both algorithms, we set $F_U^{(0)}$ is the baseline score from $I(U, d)$ in (1) after normalization such that the scores sum to 1 and $F_S^{(0)} = e^T/|V_S|$, which means we assume all speakers in the document have the equal importance.

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 [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 empirically set $\alpha = 0.9$ for all unsupervised experiments because $(1 - 0.9)$ is a proper damping factor [20, 18].

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 (Document Understanding Conference) evaluation metric, ROUGE [21], 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 well-known LexRank approach [8]. Row (c) is the result after applying random walk using topical similarity for utterance-to-utterance affinity metric. Row (d) is the result of proposed two-layer mutually reinforced random walk (MRRW) using between-layer propagation (BP)[2]. Row (e) is the result of proposed model using within- and between-layer propagation (WBP), which uses lexical similarity to measure utterance-to-utterance relation. Row (f) is the same as row (e) except it uses topical similarity for utterance-to-utterance metric.

Note that the performance of ASR is better than manual transcripts. Because a higher percentage of errors is on

[2] MRRW-BP doesn't use the similarity between utterances.

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	46.82	46.26	44.99	44.16
(b)	LexRank	48.63	48.20	46.44	45.72
(c)	Random Walk (TopicSim)	49.02	48.38	45.84	44.94
(d)	Two-Layer MRRW-BP	50.08	49.43	48.28	47.36
(e)	Two-Layer MRRW-WBP (LexSim)	49.32	48.65	**48.68**	**47.81**
(f)	Two-Layer MRRW-WBP (TopicSim)	**50.21**	**49.62**	46.82	46.19
(g)	Random Walk (TopicSim + Inter-Speaker + Intra-Speaker)	49.64	48.87	48.09	47.36
Max Relative Improvement		+7.239	+7.279	+8.211	+8.260

MRRW-BP: mutually reinforced random walk using between-layer propagation
MRRW-WBP: mutually reinforced random walk using within- and between-layer propagation

"unimportant" words, the utterances with incorrectly recognized words are harder to obtain high scores due to less similar to other utterances. Therfore, 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, 5].

3.3.1. Comparing Baseline and Single-Layer Graph Approaches

We can see the performance after basic graph-based recomputation (row (b) and row (c)) 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 graph approach is used to propagate importance based on similarity between utterances, which can effectively compensate recognition errors. Thus, sharing importance with similar utterances can significantly improve the performance for both using lexical (row (b)) and topical similarity (row (c)).

3.3.2. Comparing Single- and Two-Layer Graph Approaches

Two-layer graph approaches (row (d) - row (f)) utilize the speaker information by automatically modeling intra- and inter-speaker relation in the graph. We find that two-layer approaches involving speaker information perform better than single-layer approaches (compared to row (b) and row (c)), which means the utterances from the speakers who speak more important utterances tend to be more important [11]. Thus, propagating the importance scores between the utterances from the same speaker can improve the results. The experiment shows two-layer mutually reinforced random walk can help include the important utterances, giving better performance than traditional single-layer random walk for both ASR and manual transcripts.

3.3.3. Effectiveness of Within-Layer Propagation

Row (d) only uses between-layer propagation, while row (e) and row (f) integrate within- and between-layer propagation. For ASR transcripts, additionally considering within-layer propagation by topical similarity (row (f)) performs better than only using between-layer propagation (row (d)). For manual transcripts, within-layer propagation using lexical similarity (row (e)) improves the results from the approach via between-layer propagation (row (d)). Therefore, within-layer propagation is useful for both ASR and manual transcripts. The reason may be that two-layer mutually reinforced random walk only using between-layer propagation doesn't utilize the relation between utterances coming from different speakers. Therefore, integrating different types of relations performs better.

3.3.4. Comparing Lexical and Topical Similarity

We analyze the difference between using lexical and topical similarity for both single-layer and two-layer graph approaches. Comparing traditional LexRank and random walk with topical similarity (row (b) and row (c)), we find that ASR and manual transcripts have different results, where topical similarity performs better for ASR transcripts but worse for manual transcripts. Due to recognition errors from the recognizer, lexical similarity from word overlap may have some noises and lose some information; thus, topical similarity is better for measuring the relation between utterances from ASR output. Also, since in the absence of recognition errors lexical similarity can model the relations accurately. For two-layer graph approaches (row (e) and row (f)), they have similar condition as single-layer approaches. In conclusion, topic model helps modeling similarity between utterances from imperfect ASR transcripts for all graph-based approaches.

3.3.5. Comparison to Other Approaches

Row (g) shows the result from the approach integrating intra-speaker and inter-speaker topic modeling into a single-layer graph [11], where there are more than three parameters for controlling intra-speaker and inter-speaker topic sharing weights. However, our proposed approaches can automatically model importance of speakers and utterances by mutually propagating scores from different layers. There's only one parameter α in our unsupervised approaches. Table 1 shows our proposed approaches perform better, because they consider utterance-to-utterance, speaker-to-speaker, and speaker-to-utterance relation together by a data-driven approach. Proposed approaches achieve 7.2% and 8.2% relative improvement compared to the LTE baseline for ASR and manual transcripts respectively.

On the same corpus, Banerjee and Rudnicky [4] 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 give significant improvement, especially for ASR transcripts.

4. CONCLUSIONS AND FUTURE WORK

Extensive experiments and evaluation with ROUGE metrics showed that two-layer mutually reinforced random walk can model importance of speakers and utterances in a single graph. The speaker information can be automatically included in importance of utterances by between-layer propagation. Integrating within- and between-layer propagation can model three types of relation together, achieving about 7.2% and 8.2% relative improvement compared to the LTE baseline for ASR and manual transcripts respectively. In the future, we plan to model speakers' topic across different documents and to integrate them into a single graph.

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.

6. REFERENCES

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

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

[3] Y.-N. Chen, Y. Huang, H.-Y. Lee, and L.-S. Lee, "Spoken lecture summarization by random walk over a graph constructed with automatically extracted key terms," in *Proceedings 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 *Proceedings of IEEE SLT*, 2008.

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

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

[7] Y. Liu, S. Xie, and F. Liu, "Using N-best recognition output for extractive summarization and keyword extraction in meeting speech," in *Proceedings of IEEE ICASSP*, 2010.

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

[9] X. Cai and W. Li, "Mutually reinforced manifold-ranking based relevance propagation model for query-focused multi-document summarization," *IEEE Transactions on Acoustics, Speech and Language Processing*, vol. 20, pp. 1597–1607, 2012.

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

[11] Y.-N. Chen and F. Metze, "Integrating intra-speaker topic modeling and temporal-based inter-speaker topic modeling in random walk for improved multi-party meeting summarization," in *InterSpeech*, 2012.

[12] N. Garg, B. Favre, K. Reidhammer, and D. Hakkani-Tür, "Clusterrank: A graph based method for meeting summarization," in *Proceedings of InterSpeech*, 2009.

[13] Y.-N. Chen, C.-P. Chen, H.-Y. Lee, C.-A. Chan, and L.-S. Lee, "Improved spoken term detection with graph-based re-ranking in feature space," in *Proceedings of IEEE ICASSP*, 2011.

[14] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of SIGIR*, 1999.

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

[16] D. M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent dirichilet allocation," *Journal of Machine Learning Research*, 2003.

[17] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

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

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

[20] Brin S. and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Proceedings of WWW*, 1998.

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