Unsupervised Topic Modeling Approaches to Decision Summarization in Spoken Meetings

Lu Wang
Department of Computer Science
Cornell University
Ithaca, NY 14853
luwang@cs.cornell.edu

Claire Cardie
Department of Computer Science
Cornell University
Ithaca, NY 14853
cardie@cs.cornell.edu

Abstract

We present a token-level decision summarization framework that utilizes the latent topic structures of utterances to identify "summary-worthy" words. Concretely, a series of unsupervised topic models is explored and experimental results show that fine-grained topic models, which discover topics at the utterance-level rather than the document-level, can better identify the gist of the decision-making process. Moreover, our proposed token-level summarization approach, which is able to remove redundancies within utterances, outperforms existing utterance ranking based summarization methods. Finally, context information is also investigated to add additional relevant information to the summary.

1 Introduction

Meetings are an important way for information sharing and collaboration, where people can discuss problems and make concrete decisions. Not surprisingly, there is an increasing interest in developing methods for extractive summarization for meetings and conversations (Zechner, 2002; Maskey and Hirschberg, 2005; Galley, 2006; Lin and Chen, 2010; Murray et al., 2010a). Carenini et al. (2011) describe the specific need for *focused summaries* of meetings, i.e., summaries of a particular aspect of a meeting rather than of the meeting as a whole. For example, the decisions made, the action items that emerged and the problems arised are all important outcomes of meetings. In particular, decision summaries would allow participants to review decisions from previous meetings and understand the related topics quickly, which facilitates preparation for the upcoming meetings.

A:We decided our target group is the focus on who can afford it , (1)
B:Uh I'm kinda liking the idea of latex , if if spongy is the in thing . (2)
B:what I've seen , just not related to this , but of latex cases before , is that [vocalsound] there's uh like a hard plastic inside , and it's just covered with the latex . (2)
C:Um [disfmarker] And I think if we wanna keep our costs down , we should just go for pushbuttons , (3)
D:but if it's gonna be in a latex type thing and that's gonna look cool , then that's probably gonna have a bigger impact than the scroll wheel . (2)
A:we're gonna go with um type pushbuttons , (3)
A:So we're gonna have like a menu button , (4)
C:uh volume , favourite channels , uh and menu . (4)
A:Pre-set channels (4)

Decision Abstracts (Summary)
DECISION 1: The target group comprises of individuals who can afford the product.
DECISION 2: The remote will have a latex case.
DECISION 3: The remote will have pushbuttons.
DECISION 4: The remote will have a power button, volume buttons, channel preset buttons, and a menu button.

Figure 1: A clip of a meeting from the AMI meeting corpus (Carletta et al., 2005). A, B, C and D refer to distinct speakers; the numbers in parentheses indicate the associated meeting decision: DECISION 1, 2, 3 or 4. Also shown is the gold-standard (manual) abstract (summary) for each decision.

Meeting conversation is intrinsically different from well-written text, as meetings may not be well organized and most utterances have low density of salient content. Therefore, multiple problems need to be addressed for speech summarization. Consider the sample dialogue snippet in Figure 1 from the AMI meeting corpus (Carletta et al., 2005). Only *decision-related dialogue acts (DRDAs)* — utter-

Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 40–49,
Seoul, South Korea, 5-6 July 2012.

ances at least one decision made in the meeting[1] — are listed and ordered by time. Each DRDA is labeled numerically according to the decision it supports; so the second and third utterances (in **bold**) support DECISION 2, as do the fifth utterance in the snippet. Manually constructed *decision abstracts* for each decision are shown at the bottom of the figure.

Besides the prevalent dialogue phenomena (such as "Uh I'm kinda liking" in Figure 1), disfluencies and off-topic expressions, we notice that single utterance is usually not informative enough to form a decision. For instance, no single DRDA associated with DECISION 4 corresponds all that well with its decision abstract: "pushbuttons", "menu button" and "Pre-set channels" are mentioned in separate DAs. As a result, extractive summarization methods that select individual utterance to form the summary will perform poorly.

Furthermore, it is difficult to identify the core topic when multiple topics are discussed in one utterance. For example, all of the bold DRDAs supporting DECISION 2 contain the word "latex". However, the last DA in bold also mentions "bigger impact" and "the scroll wheel", which are not specifically relevant for DECISION 2. Though this problem can be approached by training a classifier to identify the relevant phrases and ignore the irrelevant ones or dialogue phenomena, it needs expensive human annotation and is limited to the specific domain.

Note also that for DECISION 4, the "power button" is not specified in any of the listed DRDAs supporting it. By looking at the transcript, we find "power button" mentioned in one of the preceding, but not decision-related DAs. Consequently another challenge would be to add complementary knowledge when the DRDAs cannot provide complete information.

Therefore, we need a summarization approach that is tolerant of dialogue phenomena, can determine the key semantic content and is easily transferable between domains. Recently, topic modeling approaches have been investigated and achieved state-of-the-art results in multi-document summarization (Haghighi and Vanderwende, 2009; Celikyilmaz and Hakkani-Tur, 2010). Thus, topic models appear to be a better ref for document similarity w.r.t. semantic concepts than simple literal word matching. However, very little work has investigated its role in spoken document summarization (Chen and Chen, 2008; Hazen, 2011), and much less conducted comparisons among topic modeling approaches for focused summarization in meetings.

In contrast to previous work, we study the unsupervised token-level decision summarization in meetings by identifying a concise set of key words or phrases, which can either be output as a compact summary or be a starting point to generate abstractive summaries. This paper addresses problems mentioned above and make contributions as follows:

- As a step towards creating the abstractive summaries that people prefer when dealing with spoken language (Murray et al., 2010b), we propose a token-level rather than sentence-level framework for identifying components of the summary. Experimental results show that, compared to the sentence ranking based summarization algorithms, our token-level summarization framework can better identify the summary-worthy words and remove the redundancies.
- Rather than employing supervised learning methods that rely on costly manual annotation, we explore and evaluate topic modeling approaches of different granularities for the unsupervised decision summarization at both the token-level and dialogue act-level. We investigate three topic models — Local LDA (LocalLDA) (Brody and Elhadad, 2010), Multi-grain LDA (MG-LDA) (Titov and McDonald, 2008) and Segmented Topic Model (STM) (Du et al., 2010) — which can utilize the latent topic structure on utterance level instead of document level. Under our proposed token-level summarization framework, three fine-grained models outperform the basic LDA model and two extractive baselines that select the longest and the most representative utterance for each decision, respectively. (ROUGE-SU4 F score of 14.82% for STM vs. 13.58% and 13.46% for the baselines, given the perfect clusterings of DRDAs.)
- In line with prior research that explore the role of context for utterance-based extractive summariza-

[1] These DRDAs are annotated in the AMI corpus and usually contain the decision content. They are similar, but not completely equivalent, to the *decision dialogue acts (DDAs)* of Bui et al. (2009), Fernández et al. (2008), Frampton et al. (2009).

tion (Murray and Renals, 2007), we investigate the role of context in our token-level summarization framework. For the given clusters of DRDAs, We study two types of context information — the DAs preceding and succeeding a DRDA and DAs of high TF-IDF similarity with a DRDA. We also investigate two ways to select relevant words from the context DA. Experimental results show that two types of context have comparable effect, but selecting words from the dominant topic of the center DRDA performs better than from the dominant topic of the context DA. Moreover, by leveraging context, the recall exceeds the provided upperbound's recall (ROUGE-1 recall: 48.10% vs. 45.05% for upperbound by using DRDA only) although the F scores decrease after adding context information. Finally, we show that when the true DRDA clusterings are not available, adding context can improve both the recall and F score.

2 Related Work

Speech and dialogue summarization has become important in recent years as the number of multimedia resources containing speech has grown. A primary goal for most speech summarization systems is to account for the special characteristics of dialogue. Early work in this area investigated supervised learning methods, including maximum entropy, conditional random fields (CRFs), and support vector machines (SVMs) (Buist et al., 2004; Galley, 2006; Xie et al., 2008). For unsupervised methods, maximal marginal relevance (MMR) is investigated in (Zechner, 2002) and (Xie and Liu, 2010). Gillick et al. (2009) introduce a concept-based global optimization framework by using integer linear programming (ILP).

Only in very recent works has decision summarization been addressed in (Fernández et al., 2008), (Bui et al., 2009) and (Wang and Cardie, 2011). (Fernández et al., 2008) and (Bui et al., 2009) utilize semantic parser to identify candidate phrases for decision summaries and employ SVM to rank those phrases. They also train HMM and SVM directly on a set of decision-related dialogue acts on token level and use the classifiers to identify summary-worthy words. Wang and Cardie (2011) provide an exploration on supervised and unsupervised learning for decision summarization on both utterance- and token- level.

Our work also arises out of applying topic models to text summarization (Bhandari et al., 2008; Haghighi and Vanderwende, 2009; Celikyilmaz and Hakkani-Tur, 2010; Celikyilmaz and Hakkani-Tur, 2010). Mostly, the sentences are ranked according to importance based on latent topic structures, and top ones are selected as the summary. There are some works for applying document-level topic models to speech summarization (Kong and shan Leek, 2006; Chen and Chen, 2008; Hazen, 2011). Different from their work, we further investigate the topic models of fine granularity on sentence level and leverage context information for decision summarization task.

Most existing approaches for speech summarization result in a selection of utterances from the dialogue, which cannot remove the redundancy within utterances. To eliminate the superfluous words, our work is also inspired by keyphrase extraction of meetings (Liu et al., 2009; Liu et al., 2011) and keyphrase based summarization (Riedhammer et al., 2010). However, a small set of keyphrases are not enough to concretely display the content. Instead of only picking up keyphrases, our work identifies all of the summary-worthy words and phrases, and removes redundancies within utterances.

3 Summarization Frameworks

In this section, we first present our proposed token-level decision summarization framework — **DomSum** — which utilizes latent topic structure in utterances to extract words from **Dom**inant Topic (see details in Section 3.1) to form **Sum**maries. In Section 3.2, we describe four existing sentence scoring metrics denoted as *OneTopic, MultiTopic, TMM-Sum* and *KLSum* which are also based on latent topic distributions. We adopt them to the utterance-level summarization for comparison in Section 6.

3.1 Token-level Summarization Framework

Domsum takes as input the clusters of DRDAs (with or without additional context DAs), the topic distribution for each DA and the word distribution for each topic. The output is a set of topic-coherent summary-worthy words which can be used directly as the summary or to further generate abstractive summary. We introduce DomSum in two steps according to its input: taking clusters of DRDAs as the input and with additional context information.

DRDAs Only. Given clusters of DRDAs, we use Algorithm 1 to produce the token-level summary for each cluster. Generally, Algorithm 1 chooses the topic with the highest probability as the *dominant topic* given the dialogue act (DA). Then it collects the words with a high joint probability with the dominant topic from that DA.

Input : Cluster $C = \{DA_i\}$, $P(T_j|DA_i)$, $P(w_k|T_j)$
Output: Summary

Summary$\leftarrow \Phi$ (empty set)
foreach DA_i *in* C **do**
  DomTopic$\leftarrow \max_{T_j} P(T_j|DA_i)$ (*)
  Candidate$\leftarrow \Phi$
  foreach word w_k *in* DA_i **do**
    SampleTopic$\leftarrow \max_{T_j} P(w_k|T_j)P(T_j|DA_i)$
    if DomTopic == SampleTopic **then**
      Candidate $\leftarrow$ `Union` (Candidate, w_k)
    end
  end
  Summary $\leftarrow$ `Union` (Summary, Candidate)
end

Algorithm 1: **DomSum** — The token-level summarization framework. DomSum takes as input the clusters of DRDAs and related probability distributions.

Leveraging Context. For each DRDA (denoted as "*center DA*"), we study two types of context information (denoted as "*context DAs*"). One is adjacent DAs, i.e., immediately preceding and succeeding DAs, the other is the DAs having top TF-IDF similarities with the center DA. Context DAs are added into the cluster the corresponding center DA in.

We also study two criteria of word selection from the context DAs. For each context DA, we can take the words appearing in the dominant topic of either this context DA or its center DRDA. We will show in Section 6.1 that the latter performs better as it produces more topic-coherent summaries. Algorithm 1 can be easily modified to leverage context DAs by updating the input clusters and assigning the proper dominant topic for each DA accordingly — this changes the step $(*)$ in Algorithm 1.

3.2 Utterance-level Summarization Metrics

We also adopt four sentence scoring metrics based on the latent topic structure for extractive summarization. Though they are developed on different topic models, given the desired topic distributions as input, they can rank the utterances according to their importance and provide utterance-level summaries for comparison.

OneTopic and MultiTopic. In (Bhandari et al., 2008), several sentence scoring functions are introduced based on Probabilistic Latent Semantic Indexing. We adopt two metrics, which are *OneTopic* and *MultiTopic*. For OneTopic, topic T with highest probability $P(T)$ is picked as the central topic per cluster C. The score for DA in C is:

$$P(DA|T) = \frac{\sum_{w \in DA} P(T|DA, w)}{\sum_{DA' \in C, w \in DA'} P(T|DA', w)},$$

MultiTopic modifies OneTopic by taking all of the topics into consideration. Given a cluster C, DA in C is scored as:

$$\sum_T P(DA|T)P(T) = \sum_T \frac{\sum_{w \in DA} P(T|DA, w)}{\sum_{DA' \in C, w \in DA'} P(T|DA', w)} P(T)$$

TMMSum. Chen and Chen (2008) propose a Topical Mixture Model (TMM) for speech summarization, where each dialogue act is modeled as a TMM for generating the document. TMM is shown to provide better utterance-level extractive summaries for spoken documents than other conventional unsupervised approaches, such as Vector Space Model (VSM) (Gong and Liu, 2001), Latent Semantic Analysis (LSA) (Gong and Liu, 2001) and Maximum Marginal Relevance (MMR) (Murray et al., 2005). The importance of a sentence S can be measured by its generative probability $P(D|S)$, where D is the document S belongs to. In our experiments, one decision is made per cluster of DAs. So we adopt their scoring metric to compute the generative probability of the cluster C for each DA:

$$P(C|DA) = \prod_{w_i \in C} \sum_{T_j} P(w_i|T_j)P(T_j|DA),$$

KLSum. Kullback-Lieber (KL) divergence is explored for summarization in (Haghighi and Vanderwende, 2009) and (Lin et al., 2010), where it is used to measure the distance of distributions between the document and the summary. For a cluster C of DAs, given a length limit θ, a set of DAs S is selected as:

$$S^* = \underset{S:|S|<\theta}{\arg\min}\, KL(P_C||P_S) = \underset{S:|S|<\theta}{\arg\min} \sum_{T_i} P(T_i|C) log \frac{P(T_i|C)}{P(T_i|S)}$$

4 Topic Models

In this section, we briefly describe the three fine-grained topic models employed to compute the latent topic distributions on utterance level in the

meetings. According to the input of Algorithm 1, we are interested in estimating the topic distribution for each DA $P(T|DA)$ and the word distribution for each topic $P(w|T)$. For MG-LDA, $P(T|DA)$ is computed as the expectation of local topic distributions with respect to the window distribution.

4.1 Local LDA

Local LDA (LocalLDA) (Brody and Elhadad, 2010) uses almost the same probabilistic generative model as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), except that it treats each sentence as a separate document[2]. Each DA d is generated as follows:

1. For each topic k:
   (a) Choose word distribution: $\phi_k \sim Dir(\beta)$
2. For each DA d:
   (a) Choose topic distribution: $\theta_d \sim Dir(\alpha)$
   (b) For each word w in DA d:
      i. Choose topic: $z_{d,w} \sim \theta_d$
      ii. choose word: $w \sim \phi_{z_{d,w}}$

4.2 Multi-grain LDA

Multi-grain LDA (MG-LDA) (Titov and McDonald, 2008) can model both the meeting specific topics (e.g. the design of a remote control) and various concrete aspects (e.g. the cost or the functionality). The generative process is:

1. Choose a global topic distribution: $\theta_m^{gl} \sim Dir(\alpha^{gl})$
2. For each sliding window v of size T:
   (a) Choose local topic distribution: $\theta_{m,v}^{loc} \sim Dir(\alpha^{loc})$
   (b) Choose granularity mixture: $\pi_{m,v} \sim Beta(\alpha^{mix})$
3. For each DA d:
   (a) choose window distribution: $\psi_{m,d} \sim Dir(\gamma)$
4. For each word w in DA d of meeting m:
   (a) Choose sliding window: $v_{m,w} \sim \psi_{m,d}$
   (b) Choose granularity: $r_{m,w} \sim \pi_{m,v_{m,w}}$
   (c) If $r_{m,w} = gl$, choose global topic: $z_{m,w} \sim \theta_m^{gl}$
   (d) If $r_{m,w} = loc$, choose local topic: $z_{m,w} \sim \theta_{m,v_{m,w}}^{loc}$
   (e) Choose word w from the word distribution: $\phi_{z_{m,w}}^{r_{m,w}}$

4.3 Segmented Topic Model

The last model we utilize is Segmented Topic Model (STM) (Du et al., 2010), which jointly models document- and sentence-level latent topics using a two-parameter Poisson Dirichlet Process (PDP). Given parameters α, γ, Φ and PDP parameters a, b, the generative process is:

1. Choose distribution of topics: $\theta_m \sim Dir(\alpha)$
2. For each dialogue act d:
   (a) Choose distribution of topics: $\theta_d \sim PDP(\theta_m, a, b)$
3. For each word w in dialogue act d:
   (a) Choose topic: $z_{m,w} \sim \theta_d$
   (b) Choose word: $w \sim \phi_{z_{m,w}}$

[2]For the generative process of LDA, the DAs in the same meeting make up the document, so "each DA" is changed to "each meeting" in LocalLDA's generative process.

5 Experimental Setup

The Corpus. We evaluate our approach on the AMI meeting corpus (Carletta et al., 2005) that consists of 140 multi-party meetings. The 129 scenario-driven meetings involve four participants playing different roles on a design team. A short (usually one-sentence) abstract is manually constructed to summarize each decision discussed in the meeting and used as gold-standard summaries in our experiments.

System Inputs. Our summarization system requires as input a partitioning of the DRDAs according to the decision(s) that each supports (i.e., one cluster of DRDAs per decision). As mentioned earlier, we assume for all experiments that the DRDAs for each meeting have been identified. For evaluation we consider two system input settings. In the **True Clusterings** setting, we use the AMI annotations to create perfect partitionings of the DRDAs as the input; in the **System Clusterings** setting, we employ a hierarchical agglomerative clustering algorithm used for this task in previous work (Wang and Cardie, 2011). The Wang and Cardie (2011) clustering method groups DRDAs according to their LDA topic distribution similarity. As better approaches for DRDA clustering become available, they could be employed instead.

Evaluation Metric. To evaluate the performance of various summarization approaches, we use the widely accepted ROUGE (Lin and Hovy, 2003) metrics. We use the stemming option of the ROUGE software at `http://berouge.com/` and remove stopwords from both the system and gold-standard summaries, same as Riedhammer et al. (2010) do.

Inference and Hyperparameters We use the implementation from (Lu et al., 2011) for the three topic models in Section 4. The collapsed Gibbs Sampling approach (Griffiths and Steyvers, 2004) is exploited for inference. Hyperparameters are chosen according to (Brody and Elhadad, 2010), (Titov and McDonald, 2008) and (Du et al., 2010). In LDA and LocalLDA, α and β are both set to 0.1 . For MG-LDA, α^{gl}, α^{loc} and α^{mix} are set to 0.1; γ is 0.1

and the window size T is 3. And the number of local topic is set as the same number of global topic as discussed in (Titov and McDonald, 2008). In STM, α, a and b are set to 0.5, 0.1 and 1, respectively.

5.1 Baselines and Comparisons

We compare our token-level summarization framework based on the fine-grained topic models to (1) two unsupervised baselines, (2) token-level summarization by LDA, (3) utterance-level summarization by Topical Mixture Model (TMM) (Chen and Chen, 2008), (4) utterance-level summarization based on the fine-grained topic models using existing metrics (Section 3.2), (5) two supervised methods, and (6) an upperbound derived from the AMI gold standard decision abstracts. (1) and (6) are described below, others will be discussed in Section 6.

The LONGEST DA Baseline. As in (Riedhammer et al., 2010) and (Wang and Cardie, 2011), this baseline simply selects the longest DRDA in each cluster as the summary. Thus, it performs utterance-level decision summarization. This baseline and the next allow us to determine summary quality when summaries are restricted to a single utterance.

The PROTOTYPE DA Baseline. Following Wang and Cardie (2011), the second baseline selects the decision cluster prototype (i.e., the DRDA with the largest TF-IDF similarity with the cluster centroid) as the summary.

Upperbound. We also compute an upperbound that reflects the gap between the best possible extractive summaries and the human-written abstracts according to the ROUGE score: for each cluster of DRDAs, we select the words that also appear in the associated decision abstract.

6 Results and Discussion

6.1 True Clusterings

How do fine-grained topic models compare to basic topic models or baselines? Figure 2 demonstrates that by using the DomSum token-level summarization framework, the three fine-grained topic models uniformly outperform the two non-trivial baselines and TMM (Chen and Chen, 2008) (reimplemented by us) that generates utterance-level summaries. Moreover, the fine-grained models also beat basic LDA under the same DomSum token-level summarization framework. This shows the fine-grained topic models that discover topic structures on utterance-level better identify gist information.

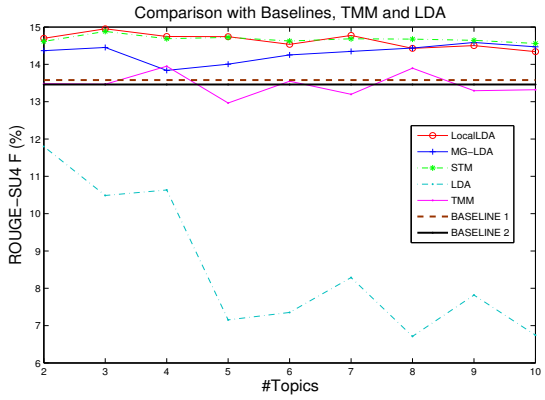

Figure 2: With true clusterings of DRDAs as the input, we use DomSum to compare the performance of LocalLDA, MGLDA and STM against two baselines, LDA and TMM. "# topic" indicates the number of topics for the model. For MGLDA, "# topic" is the number of local topics.

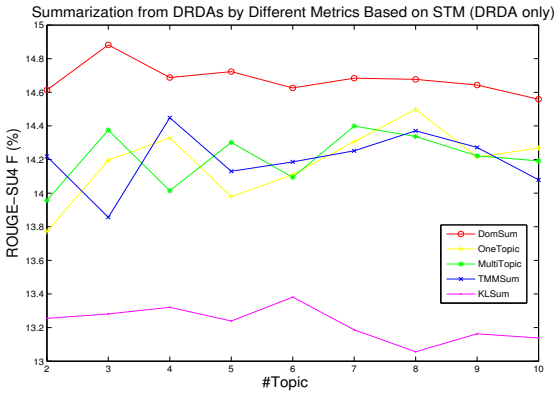

Figure 3: With true clusterings of DRDAs as the input, DomSum is compared with four DA-level summarization metrics using topic distributions from STM. Results from LocalLDA and MGLDA are similar so they are not displayed.

Can the proposed token-level summarization framework better identify important words and remove redundancies than utterance selection methods? Figure 3 demonstrates the comparison results for our DomSum token-level summarization framework with four existing utterance scoring metrics discussed in Section 3.2, namely OneTopic, MultiTopic, TMMSum and KLSum. The utterance with highest score is extracted to form the summary. LocalLDA and STM are utilized to compute the input distributions, i.e., $P(T|DA)$ and $P(w|T)$. From Figure 3, DomSum yields the best F scores which

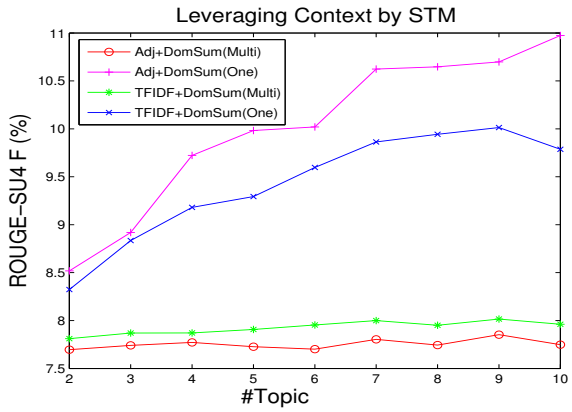

Figure 4: Under DomSum framework, two types of context information are added: Adjacent DA ("Adj") and DAs with high TFIDF similarities ("TFIDF"). For each context DA, selecting words from the dominant topic of center DA ("One") or the current context DA ("Multi") are investigated.

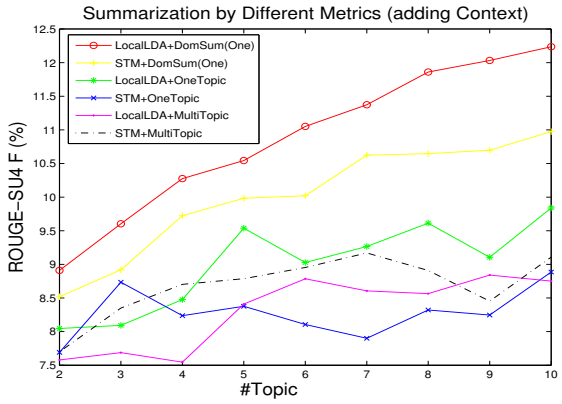

Figure 5: By using adjacent DAs as context, DomSum is compared with two DA-level summarization metrics: OneTopic and MultiTopic. For DomSum, the words of context DA from dominant topic of the center DA ("One") is selected; For OneTopic and MultiTopic, three top ranked DAs are selected.

shows that the token-level summarization approach is more effective than utterance-level methods.

Which way is better for leveraging context information? We explore two types of context information. For adjacent content (*Adj* in Figure 4), 5 DAs immediately preceding and 5 DAs succeeding the center DRDA are selected. For TF-IDF context (*TFIDF* in Figure 4), 10 DAs of highest TF-IDF similarity with the center DRDA are taken. We also explore two ways to extract summary-worthy words from the context DA — selecting words from the dominant topic of either the center DA (denoted as "One" in parentheses in Figure 4) or the current context DA (denoted as "multi" in parentheses in Figure 4). Figure 4 indicates that the two types of context information do not have significant difference, while selecting the words from the dominant topic of the center DA results in better ROUGE-SU4 F scores. Notice that compared with Figure 3, the results in Figure 4 have lower F scores when using the true clusterings of DRDAs. This is because context DAs bring in relevant words as well as noisy information. We will show in Section 6.2 that when true clusterings are not available, the context information can boost both recall and F score.

	True Clusterings				
		R-1		**R-2**	**R-SU4**
	PREC	REC	F1	F1	F1
Baselines					
Longest DA	34.06	31.28	32.61	12.03	13.58
Prototype DA	40.72	28.21	33.32	12.18	13.46
Supervised Methods					
CRF	52.89	26.77	35.53	11.48	14.03
SVM	43.24	37.92	40.39	12.78	16.24
Our Approach					
5 topics					
LocalLDA	35.18	38.92	36.95	12.33	14.74
+ *context*	17.26	45.34	25.00	8.40	11.05
STM	34.06	41.30	**37.32**	**12.42**	**14.82**
+ *context*	15.60	***48.10***	23.56	8.16	9.98
10 topics					
LocalLDA	36.20	36.81	36.50	12.04	14.34
+ *context*	21.82	41.57	28.62	9.61	12.24
STM	34.15	40.83	37.19	12.40	14.56
+ *context*	17.87	46.57	25.82	8.89	10.97
Upperbound	100.00	***45.05***	62.12	33.27	34.89

Table 1: ROUGE-1 (R-1), ROUGE-2 (R-2) and ROUGE-SU4 (R-SU4) scores for our proposed token-level summarization approaches along with two baselines, supervised methods and the Upperbound (only using DRDAs). — all use True Clusterings

How do the token-level summarization framework compared to utterance selection methods for leveraging context? We also compare the ability of leveraging context of DomSum to utterance scoring metrics, i.e., OneTopic and MultiTopic. 5 DAs preceding and 5 DAs succeeding the center DA are added as context information. For context DA under DomSum, we select words from the dominant topic of the center DA (denoted as "One" in parentheses in Figure 5). For OneTopic and MultiTopic, the top 3 DAs are extracted as the summary. Figure 5 demonstrates the combination of LocalLDA and STM with each of the metrics. DomSum, as a token-level summarization metrics, dominates other two metrics in leveraging context.

	System Clusterings				
	R-1			R-2	R-SU4
	PREC	REC	F1	F1	F1
Baselines					
Longest DA	17.06	11.64	13.84	2.76	3.34
Prototype DA	18.14	10.11	12.98	2.84	3.09
Supervised Methods					
CRF	46.97	15.25	23.02	6.09	9.11
SVM	39.05	18.45	25.06	6.11	9.82
Our Approach					
5 topics					
LocalLDA	25.57	16.57	20.11	4.03	5.87
+ *context*	20.68	25.96	23.02	3.09	4.48
STM	24.15	17.82	20.51	4.03	5.69
+ *context*	20.64	***30.03***	**24.47**	3.59	4.76
10 topics					
LocalLDA	25.98	15.94	19.76	3.59	4.41
+ *context*	23.98	21.92	22.90	3.45	4.10
STM	26.32	19.14	22.16	**4.07**	**5.88**
+ *context*	22.50	28.40	25.11	3.43	4.15

Table 2: ROUGE-1 (R-1), ROUGE-2 (R-2) and ROUGE-SU4 (R-SU4) scores for our proposed token-level summarization approaches, compared with two baselines and supervised methods. — all use System Clusterings

DRDA (1): I think if we can if we can include them at not too much extra cost, then I'd put them in, **DRDA (2)**: Uh um we we're definitely going in for voice recognition as well as LCDs, mm. **DRDA (3)**: So we've basically worked out that we're going with a simple battery, **context DA (1)**:So it's advanced integrated circuits? **context DA (2)**:the advanced chip **context DA (3)**: and a curved on one side case which is folded in on itself , um made out of rubber
Decision Abstract: It will have voice recognition, use a simple battery, and contain an advanced chip.
Longest DA & Prototype DA: Uh um we we're definitely going in for voice recognition as well as LCDs, mm. **TMM:** I think if we can if we can include them at not too much extra cost, then I'd put them in, **SVM:** cost voice recognition simple battery **CRF:** voice recognition battery **STM:** extra cost, definitely going voice recognition LCDs, simple battery **STM + context:** cost, company, advanced integrated circuits, going voice recognition, simple battery, advanced chip, curved case rubber

Table 3: Sample system outputs by different methods are in the third cell (methods' names are in bold). First cell contains three DRDAs supporting the decision in the second cell and three adjacent DAs of them.

How do our approach perform when compared with supervised learning approaches? For a better comparison, we also provide summarization results by using supervised systems along with an upperbound. We use Support Vector Machines (Joachims, 1998) with RBF kernel and order-1 Conditional Random Fields (Lafferty et al., 2001) — trained with the same features as (Wang and Cardie, 2011) to identify the summary-worthy **tokens** to include in the abstract. A three-fold cross validation is conducted for both methods. ROUGE-1, ROUGE-2 and ROUGE-SU4 scores are listed in Table 1. From Table 1, our token-level summarization approaches based on LocalLDA and STM are shown to outperform the baselines and even the CRF. Meanwhile, by adding context information, both LocalLDA and STM can get better ROUGE-1 recall than the supervised methods, even higher than the provided upperbound which is computed by only using DRDAs. This shows the DomSum framework can leverage context to compensate the summaries.

6.2 System Clusterings

Results using the **System Clusterings** (Table 2) present similar findings, though all of the system and baseline scores are lower. By adding context information, the token-level summarization approaches based on fine-grained topic models compare favorably to the supervised methods in F scores, and also get the best ROUGE-1 recalls.

6.3 Sample System Summaries

To better exemplify the summaries generated by different systems, sample output for each method is shown in Table 3. We see from the table that utterance-level extractive summaries (Longest DA, Prototype DA, TMM) make more coherent but still far from concise and compact abstracts. On the other hand, the supervised methods (SVM, CRF) that produce token-level extracts better identify the overall content of the decision abstract. Unfortunately, they require human annotation in the training phase. In comparison, the output of fine-grained topic models can cover the most useful information.

7 Conclusion

We propose a token-level summarization framework based on topic models and show that modeling topic structure at the utterance-level is better at identifying relevant words and phrases than document-level models. The role of context is also studied and shown to be able to identify additional summary-worthy words.

Acknowledgments This work was supported in part by National Science Foundation Grants IIS-0968450 and IIS-1111176, and by a gift from Google.

References

Harendra Bhandari, Takahiko Ito, Masashi Shimbo, and Yuji Matsumoto. 2008. Generic text summarization using probabilistic latent semantic indexing. In *Proceedings of IJCNLP*, pages 133–140.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 804–812, Stroudsburg, PA, USA. Association for Computational Linguistics.

Trung H. Bui, Matthew Frampton, John Dowding, and Stanley Peters. 2009. Extracting decisions from multi-party dialogue using directed graphical models and semantic similarity. In *Proceedings of the SIGDIAL 2009 Conference*, pages 235–243.

Anne Hendrik Buist, Wessel Kraaij, and Stephan Raaijmakers. 2004. Automatic summarization of meeting data: A feasibility study. In *Proc. Meeting of Computational Linguistics in the Netherlands (CLIN)*.

Giuseppe Carenini, Gabriel Murray, and Raymond Ng. 2011. *Methods for Mining and Summarizing Text Conversations*. Morgan & Claypool Publishers.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, and Mccowan Wilfried Post Dennis Reidsma. 2005. The ami meeting corpus: A pre-announcement. In *In Proc. MLMI*, pages 28–39.

Asli Celikyilmaz and Dilek Hakkani-Tur. 2010. A hybrid hierarchical model for multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 815–824, Stroudsburg, PA, USA. Association for Computational Linguistics.

Berlin Chen and Yi-Ting Chen. 2008. Extractive spoken document summarization for information retrieval. *Pattern Recogn. Lett.*, 29:426–437, March.

Lan Du, Wray Buntine, and Huidong Jin. 2010. A segmented topic model based on the two-parameter poisson-dirichlet process. *Mach. Learn.*, 81:5–19, October.

Raquel Fernández, Matthew Frampton, John Dowding, Anish Adukuzhiyil, Patrick Ehlen, and Stanley Peters. 2008. Identifying relevant phrases to summarize decisions in spoken meetings. INTERSPEECH-2008, pages 78–81.

Matthew Frampton, Jia Huang, Trung Huu Bui, and Stanley Peters. 2009. Real-time decision detection in multi-party dialogue. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, pages 1133–1141.

Michel Galley. 2006. A skip-chain conditional random field for ranking meeting utterances by importance. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 364–372.

Dan Gillick, Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tur. 2009. A global optimization framework for meeting summarization. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '09, pages 4769–4772. IEEE Computer Society.

Yihong Gong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 19–25, New York, NY, USA. ACM.

T. L. Griffiths and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 362–370. Association for Computational Linguistics.

Timothy J. Hazen. 2011. Latent topic modeling for audio corpus summarization. In *INTERSPEECH*, pages 913–916.

Thorsten Joachims. 1998. Text categorization with Support Vector Machines: Learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Machine Learning: ECML-98*, volume 1398, chapter 19, pages 137–142. Berlin/Heidelberg.

Sheng-Yi Kong and Lin shan Leek. 2006. Improved spoken document summarization using probabilistic latent semantic analysis (plsa). In *Proceedings of the 2006 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '06.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Shih-Hsiang Lin and Berlin Chen. 2010. A risk minimization framework for extractive speech summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 79–87. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, pages 71–78.

S.-H. Lin, Y.-M. Yeh, and B. Chen. 2010. Leveraging kullback-leibler divergence measures and information-rich cues for speech summarization.

Feifan Liu, Deana Pennell, Fei Liu, and Yang Liu. 2009. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 620–628, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fei Liu, Feifan Liu, and Yang Liu. 2011. A supervised framework for keyword extraction from meeting transcripts. *IEEE Transactions on Audio, Speech & Language Processing*, 19(3):538–548.

Bin Lu, Myle Ott, Claire Cardie, and Benjamin Tsou. 2011. Multi-aspect sentiment analysis with topic models. In *Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction*.

Sameer Maskey and Julia Hirschberg. 2005. Comparing Lexical, Acoustic/Prosodic, Structural and Discourse Features for Speech Summarization. In *Proc. European Conference on Speech Communication and Technology (Eurospeech)*.

Gabriel Murray and Steve Renals. 2007. Towards online speech summarization. In *INTERSPEECH*, pages 2785–2788.

Gabriel Murray, Steve Renals, and Jean Carletta. 2005. Extractive summarization of meeting recordings. In *in Proceedings of the 9th European Conference on Speech Communication and Technology*, pages 593–596.

Gabriel Murray, Giuseppe Carenini, and Raymond Ng. 2010a. Interpretation and transformation for abstracting conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 894–902, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gabriel Murray, Giuseppe Carenini, and Raymond T. Ng. 2010b. Generating and validating abstracts of meeting conversations: a user study. In *INLG'10*.

Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tür. 2010. Long story short - global unsupervised models for keyphrase based meeting summarization. *Speech Commun.*, 52(10):801–815, October.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 111–120. ACM.

Lu Wang and Claire Cardie. 2011. Summarizing decisions in spoken meetings. In *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, pages 16–24, Portland, Oregon, June. Association for Computational Linguistics.

Shasha Xie and Yang Liu. 2010. Using confusion networks for speech summarization. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 46–54. Association for Computational Linguistics.

Shasha Xie, Yang Liu, and Hui Lin. 2008. Evaluating the effectiveness of features and sampling in extractive meeting summarization. In *Proc. of IEEE Spoken Language Technology (SLT)*.

Klaus Zechner. 2002. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Comput. Linguist.*, 28:447–485, December.