Summarizing Decisions in Spoken Meetings

Lu Wang and Claire Cardie

Department of Computer Science, Cornell University

Cornell University

Problem Description

Input: A Clip of a Meeting
C: Just spinning and not scrolling , I would say . (1)
C: But if you've got a [disfmarker] if if you've got a flipped thing , effectively it's something that's curved on one side and flat on the other side , but you folded it in half . (2)
D: the case would be rubber and the the buttons , (3)
B: I think the spinning wheel is definitely very now . (1)
B: and then make the colour of the main remote [vocalsound] the colour like vegetable colours , do you know ? (4)
B: I mean I suppose vegetable colours would be orange and green and some reds and um maybe purple (4)
A: but since LCDs seems to be uh a definite yes , (1)
A: Flat on the top . (2)
Output: Decision Abstracts (Summary)
DECISION 1: The remote will have an LCD and spinning wheel inside.
DECISION 2: The case will be flat on top and curved on the bottom.
DECISION 3: The remote control and its buttons will be made of rubber.
DECISION 4: The remote will resemble a vegetable and be in bright vegetable colors.

Table 1: 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. This table lists only ***decision-related dialogue acts (DRDAs)*** — utterances associated with at least one decision made in the meeting.

Challenges:

- The utterances associated with a single decision are not contiguous in the dialogue. For example, the dialogue acts concerning DECISION 1 are interleaved with DAs for other decisions.
- Some decision-related DAs contribute more than others to the associated decision. In composing the summary for DECISION 1, for example, we might safely ignore the first DA for DECISION 1.
- More so than for standard text summarization, purely extract-based summaries are not likely to be easily interpretable: DRDAs often contain text that is irrelevant to the decision and many will only be understandable if analyzed in the context of the surrounding utterances.

Clustering Decision-Related Dialogue Acts

General Framework: Hierarchical Agglomerative Clustering

Approach1: Unsupervised Clustering

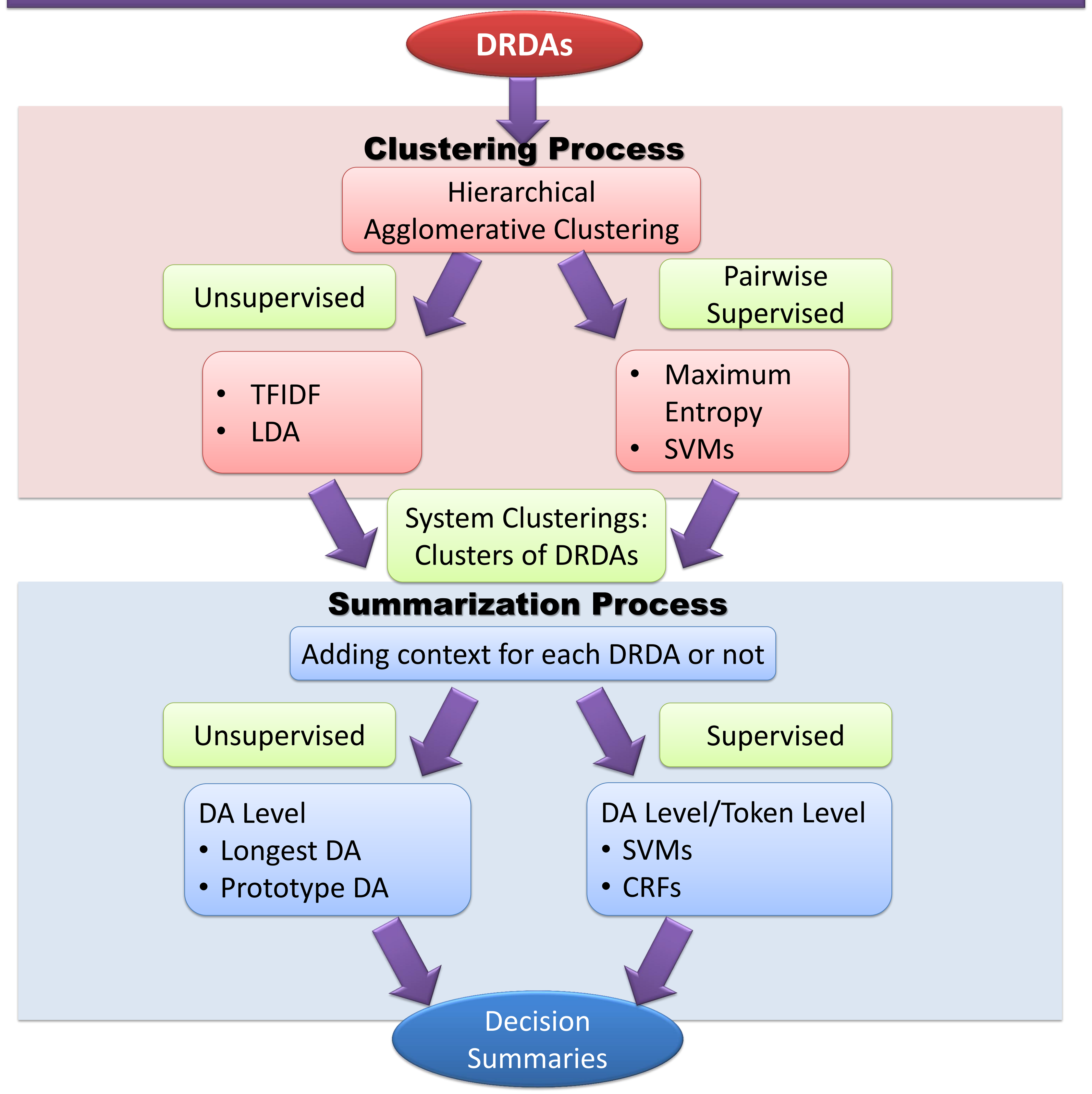
- Each DRDA is represented as a feature vector $\overrightarrow{FV} = (x_1, x_2, \ldots, x_n)$,
  - TFIDF similarity: x_i is word w_i's TFIDF score
  - LDA: x_i is topic T_i's probability

Approach2: Pairwise Supervised Clustering

- Use a classifier to determine whether pairwise DAs should be in the same cluster
- Each feature vector is derived from a pair of DAs

Table 2: Features for Clustering
number of overlapping words
proportion of the number of overlapping words to the length of shorter DA
TF-IDF similarity
whether the DAs are in an adjacency pair
time difference of pairwise DAs
relative dialogue position of pairwise DAs
whether the two DAs have the same DA type
number of overlapping words in the contexts

Summarization Framework



Decision Summarization

Approach1: Unsupervised Methods for DA Level Summarization

- Longest DA in the cluster
- Prototype DA: the DRDA with the largest TFIDF similarity with the cluster centroid

Approach2: Supervised Methods for DA/Token Level Summarization

Table 3: Features for DA Level Summarization
Lexical Features
unigram/bigram
length of the DA
contain digits?
has overlapping words with next DA?
next DA is a positive feedback?
Structural Features
relative position in the meeting?
in an AP?
if in an AP, AP type
if in an AP, the other part is decision-related?
if in an AP, is the source part or target part?
if in an AP and is source part, target is positive feedback?
if in an AP and is target part, source is a question?
Discourse Features
relative position to "WRAP UP" or "RECAP"
Other Features
DA type/speaker role/topic

Table 4: Features for Token Level Summarization
Lexical Features
current token/current token and next token length of the DA
is digit?
appearing in next DA?
next DA is a positive feedback?
Structural Features
see Table 3
Grammatical Features
part-of-speech
phrase type (VP/NP/PP)
dependency relations
Other Features
speaker role/topic

Results

SUMMARIZATION / CLUSTERING		via CRFs			via SVMs		
		PRECISION	RECALL	F1	PRECISION	RECALL	F1
True Clusterings							
	DA	0.3922	0.4449	**0.3789**	0.3661	0.4695	0.3727
	Token	0.5055	0.2453	0.3033	0.4953	0.3788	**0.3963**
	DA+Context	0.3753	0.4372	**0.3678**	0.3595	0.4449	0.3640
	Token+Context	0.5682	0.2825	0.3454	0.6213	0.3868	**0.4387**
System Clusterings							
	using LDA						
	DA	0.3087	0.1663	**0.1935**	0.3391	0.2097	**0.2349**
	Token	0.3379	0.0911	0.1307	0.3760	0.1427	0.1843
	DA+Context	0.3305	0.1748	**0.2041**	0.2903	0.1869	**0.2068**
	Token+Context	0.4557	0.1198	0.1727	0.4882	0.1486	0.2056
	using SVMs						
	DA	0.3508	0.1884	**0.2197**	0.3592	0.2026	**0.2348**
	Token	0.2807	0.0497	0.0777	0.3607	0.0885	0.1246
	DA+Context	0.3583	0.1891	**0.2221**	0.3418	0.1892	**0.2213**
	Token+Context	0.4891	0.0822	0.1288	0.4873	0.0914	0.1393
No Clustering							
	DA	0.0867	0.1957	0.0993	0.0707	0.1979	0.0916
	Token	0.1906	0.0625	0.0868	0.1890	0.3068	**0.2057**

Table 5: Supervised Summarization Results by using True/System/No Clustering

SUMMARIZATION / CLUSTERING		PRECISION	RECALL	F1
True Clusterings				
	Longest DA	0.3655	0.4077	**0.3545**
	Prototype DA	0.3626	0.4140	0.3539
System Clusterings				
	using LDA			
	Longest DA	0.3623	0.1892	**0.2214**
	Prototype DA	0.3669	0.1887	0.2212
	using SVMs			
	Longest DA	0.3719	0.1261	0.1682
	Prototype DA	0.3816	0.1264	0.1700
No Clustering				
	Longest DA	0.1039	0.1382	0.1080
	Prototype DA	0.1350	0.1209	0.1138
Upper Bound		0.8970	0.4089	**0.5333**

Table 6: Unsupervised Summarization Results by using True/System/No Clustering

Conclusion

- Among the unsupervised clustering methods, the LDA topic modeling is preferred to TFIDF. For the pairwise supervised clustering methods, SVMs and Maximum Entropy produce comparable results.
- SVMs have a superior or comparable summarization performance vs. CRFs on every task.
- Token-level summaries perform better than DA-level summaries only using TRUE CLUSTERINGS and the SVM-based summarizer.
- Discourse context generally improves token-level summaries but not DA-level summaries.
- Clusterings produced by (unsupervised) LDA lead to summaries that are quite comparable in quality to those generated from DRDA clusterings produced by SVMs (supervised).