Finding Topics in Emails: Are LDA & LCSeg enough?

Shafiq Joty, Giuseppe Carenini, Gabriel Murray, and Raymond Ng

An email thread often discusses multiple topics

•E.g., email thread about arranging **a conference** may cover:

Location and time | Registration

Food menu | Workshops

The Finding Topics Task:

➢Clustering the sentences into a set of coherent clusters.

Prerequisite for:

➢Higher-level conversation analysis (e.g., speech act tagging).

➢Text summarization & Automatic question answering.

➢Information retrieval.

Evaluation Metrics:

•**Kappa statistics is not applicable:** 2 annotations may have different number of topics.

•**We adopt the metrics used in [1] for chats:**

- **1-to-1** measures the global similarity by pairing up the clusters of 2 annotations to maximize the total overlap.
- **loc_k** measures the local agreement within a context of k utterances.
- **m-to-1** measures how much two annotators agree on the general structure by mapping each of the clusters of the 1st annotation to the single cluster in the 2nd annotation with which it has the greatest overlap.

Our Preliminary Development Set

➢**5** email threads, **avg. 3.5 topics**

➢**Annotator Agreement of 4 Human Annotators:**

	Chat Corpus [1]			Our E-mail Corpus		
Scores	Mean	Max	Min	Mean	Max	Min
1-to-1	52.98	63.50	35.63	64.99	100	39.13
loc_3	81.09	86.53	74.75	69.50	100	40
m-to-1 (by entropy)	86.70	94.13	75.50	85.42	100	65.22

Table 1: annotator agreement: chat corpus and our email corpus

Existing Methods: LDA & LCSeg

How LDA can be used?

•LDA gives: distribution of words over the topics, **$P(z_i = j \mid w_i)$.**

•Assuming the words in a sentence occur independently, the distribution of sentences is calculated as follows:

$$P(z_i = j|s_k) = \prod_{w_i \in s_k} P(z_i = j|w_i)$$

•Topic assignment is done by:

$$j^* = argmax_j P(z_i = j|s_k)$$

Problems with LDA

•LDA based models are:

➢based on only lexical distribution.

➢inadequate especially when topics are closely related.

•Does not consider email specific features:

➢conversational structure

➢mentioning each other's name,

➢cue phrases, subject of the email, sender.

Lexical Chain based Approach: LCSeg

• Emails are ordered based on the **temporal relation**.

• Computes lexical chains based on word repetition.

• Start and end of strong repetitions indicate topic boundary.

Problems with LCSeg

•Considers only:

➢ minimal conversation structure (temporal relation between emails)

➢ lexical cohesion

•Does not consider:

➢ Fine-grained conversation structure

➢ global similarity between sentences

➢ other important features (e.g., cue phrases, sender, subject, etc.)

Fine-grained conversation structure

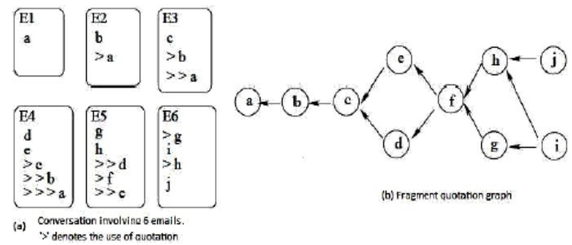

Figure 1: Fragment Quotation Graph for emails

Proposed Solution

• Capture the conversation structure using Fragment Quotation Graph (FQG)

• Consider a rich feature set (for each pair of sentences):

- **Topic features:** LSA, LDA.
- **Conversation features:** distance between two sentences in FQG, speaker, mention of names.
- **Lexical features:** tf*idf, Cue words.

• Inspired by [1] a binary classifier is learned to decide, given any two sentences, whether they should be in the same topic or not.

• Form an undirected graph G = (V,E),

- V represents the sentences
- Edge weights w (u,v) denote the class membership probability

•The problem then becomes **graph partitioning problem** which we solve using the **Normalized Cut criteria**.

Preliminary Results

Metrics	LDA			LCSeg			Proposed Solution
	Max.	Avg.	Min.	Max.	Avg.	Min.	……..
1-to-1	0.67	0.49	0.30	0.8	0.54	0.35	……..
Loc3	0.74	0.52	0.33	0.83	0.62	0.42	……..
m-to-1	0.81	0.68	0.6	1.0	0.75	0.59	……..

Reference:

[1] M. Elsner and E. Charniak. You talking to me? a corpus and algorithm for conversation disentanglement. In *Proceedings of ACL-08*. ACL.