Finding Topics in Emails: Is LDA enough?

Shafiq Joty, Giuseppe Carenini, Gabriel Murray, and Raymond Ng
Department of Computer Science
University of British Columbia
Vancouver, B.C. Canada V6T1Z4
{rjoty,carenini,gabrielm,rng}@cs.ubc.ca

Abstract

Our research addresses the task of finding topics at the sentence level in email conversations. As an asynchronous collaborative application, email has its own characteristics which differ from written monologues (e.g., text books, news articles) or spoken dialogs (e.g., meetings). Hence, the generative topic models like Latent Dirichlet Allocation (LDA) and its variations, which are successful in finding topics in monologue or dialog, may not be successful by themselves in asynchronous written conversations like emails. However, an effective combination of LDA with other important features can give us the desired results. We first point out the specific characteristics of emails that we need to consider in order to find the inherent topics discussed in an email conversation. Then we demonstrate why the generative topic models by themselves may not be adequate for this task. We propose a novel graph-theoretic framework to solve the problem. Crucial to our proposed approach is that it captures the discriminative email features and integrates the strengths of the supervised approach with the unsupervised technique considering LDA yet as one of the important factors.

1 Introduction

Our definition of 'topic' is something about which the participants of a conversation discuss or argue. For example, an email thread about arranging a conference can have topics such as 'location and time', 'registration', 'food menu', 'workshops', etc. Multiple topics seem to occur naturally in social interactions, whether synchronous (e.g., chats, meetings) or asynchronous (e.g., emails, blogs) conversations. In multi-party chat [5] report an average of 2.75 discussions. In our current, still limited development set containing 5 email threads from the BC3 corpus[1], we found an average of 3.5 topics per thread.

Finding topics in our approach involves clustering the sentences of a thread into a set of coherent clusters. It is often considered as a prerequisite for other higher-level conversation analysis (i.e., identifying dialog acts, adjacency pairs, and rhetorical relations) and the applications of the derived structure are broad, encompassing text summarization, information ordering, automatic question answering, information retrieval and intelligent user interfaces.

To our knowledge there is no previous research that tries to find topics at the sentence level in emails, though the closely related task of "topic segmentation" in monologue and dialog has received extensive attention. Topic segmentation or finding topic boundaries is the task of partitioning a text into a linear sequence of topically coherent segments. Previous work in topic segmentation can be classified into two broad classes: unsupervised and supervised. All unsupervised approaches proposed so far, have been applied to capture only a few

[1] http://www.cs.ubc.ca/nest/lci/bc3.html

factors/features related to the topic (e.g., lexical distribution, author), and these factors may not generalize well across multiple datasets. On the other hand, supervised approaches perform better than the unsupervised techniques ([6]) as they can easily incorporate many informative features. However, they require a huge amount of manually labeled data. Any approach that integrates the benefits of the supervised approach (i.e., being able to incorporate several important features) with the benefits of unsupervised approach (i.e., not requiring a huge amount manually annotated data) is highly desirable.

The probabilistic topic models (e.g., [14]) have proven to be successful for topic segmentation in both monologue (e.g., [4]) and dialog (e.g., [7]). However, email has its own characteristics and to extract the topic structure from an email thread successfully a method must consider these characteristics. In this paper, we argue that probabilistic topic models by themselves (e.g., LDA), which are mostly based on lexical distribution, are not capable of considering all these important email features. However, an effective combination of LDA with these features can give us the desired results. Therefore, we propose a novel graph-theoretic framework that captures the discriminative email features and integrates the strengths of supervised approach with the unsupervised technique keeping LDA as one of the crucial components.

In Section 2 we first discuss the evaluation metrics we are using to develop and test our approach. In section 3 we show how we can use the probabilistic topic models in an initial attempt to extract the topic structure from an email thread. Then we point out the email specific features that one needs to consider and argue why LDA is not enough to serve our purpose. In Section 4 we propose our solution to remedy these problems.

2 Evaluation Metrics

We want to compare two annotations (or systems' output) where the number of topic clusters in these annotations (or systems' output) may differ. As the κ statistic is not applicable here [5], we adapt the metrics used in [5] to measure both the inter-annotator agreement and the systems' performance. Specifically, to measure global similarity between annotations, we use *1-to-1* measure that pairs up the clusters from the two annotations to maximize the total overlap and reports the percentage of overlap. To measure local agreement we use loc_k that measures the agreement between two annotations within a context of k utterances. For example, loc_1 measures the pairs of adjacent utterances for which two annotations agree. To measure how much two annotators agree on the general structure, we apply entropy based metric $m - to - 1$ that maps each of the clusters of the first annotation to the single cluster in the 2nd annotation with which it has the greatest overlap, then counts the percentage of overlap. Table 1 shows that annotator agreement (4 annotators) on our email corpus is similar to the one (6 annotators) for the more extensive chat corpus used in [5].

	Chat Corpus [5]			Our 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

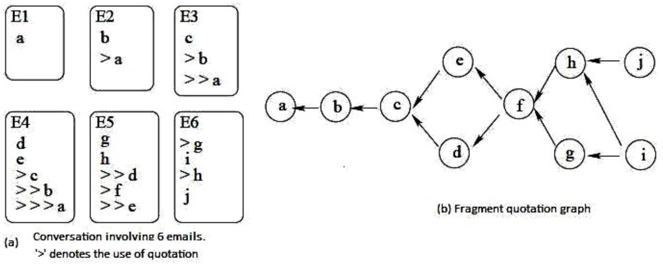
Table 1: annotator agreement: chat corpus and email corpus

3 Is LDA Enough?

We first show the way probabilistic topic models can be formalized to solve the problem of finding topics in an email thread. These models rely on the same fundamental idea: documents are mixtures of topics, where a topic is a probability distribution over words [14]. To produce a new document, at first we choose a distribution over topics. Then, for each word in that document, we choose a topic at random according to this distribution, and draw a word from that topic. The generative topic model specifies the following distribution over words within a document: $P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j)$, where T is the number of topics. $P(w_i|z_i = j)$ is the probability of word w_i under topic j and $P(z_i = j)$ is the probability that j^{th} topic was sampled for the i^{th} word token. We refer the multinomial distributions $\phi^{(j)} = P(w|z_i = j)$ and $\theta^{(d)} = P(z)$ as topic-word distribution and document-topic distribution respectively. [1] refined this basic model by placing a Dirichlet (α) prior on θ. [8] further refined it by placing a Dirichlet (β) prior on ϕ. Now, the inference problem is to find ϕ and θ given a document set. EM has been applied to estimate these two parameters directly. Instead of estimating ϕ and θ, we can also directly estimate the posterior distribution over $z = P(z_i = j|w_i)$ (topic assignments for words). One efficient estimation technique uses Gibbs sampling to estimate this distribution.

This framework which makes the bag-of-words (BOW) assumption, can be directly applied to an email corpus by considering each email as a document. So, all the emails in the email corpus (note, not a single thread) constitute the document set. Using LDA we get $z = P(z_i = j|w_i)$ (i.e., topic assignments for words). By assuming the words in a sentence occur independently we can estimate the the topic assignments for sentences as follows: $P(z_i = j|s_k) = \prod_{w_i \in s_k} P(z_i = j|w_i)$ where, s_k is the k^{th} sentence for which we can assign the topic by: $j^* = argmax_j P(z_i = j|s_k)$.

Although several improvements of LDA over the BOW approach have been proposed (e.g., [12], [2], [9]), we argue that LDAs are still inadequate for finding topics in emails especially when topics are closely related (e.g., 'extending the meeting' and 'scheduling the meeting') and distributional variations are subtle. To better identify the topics in an email thread we need to consider the email specific features. The most important features is the 'conversation structure'. We [3] previously showed how this structure can be captured efficiently at the finer granularity level (i.e., fragment level) using the 'fragment quotation graph'. Based on an analysis of the quotation embedded in emails, the graph provides a fine representation of the referential structure of a conversation. Figure 1 shows one example of 6 emails in an email thread (a) and their corresponding fragment quotation graph (b). Note that the fragment quotation graph can also handle the hidden email problem [3]. In our development set we found people often use quotation to refer to the same topic. Another very important feature in emails and in multi-party chat is 'mentioning names'. [11] hypothesize that mentioning each other's name is a strategy that participants use in multi-party written conversations to make disentanglement easier. When people reply to multiple recipients they usually mention the name of the person being referred to. Another key feature in any discourse is 'topic shift cue words' like "now", "however", etc. that people often use to shift from one topic to another.



(a) Conversation involving 6 emails. '>' denotes the use of quotation

(b) Fragment quotation graph

Figure 1: Fragment Quotation Graph for emails

Our hypothesis is that LDA is not capable of capturing these features. However, combining LDA with other conversational features can be very effective. In the next section we propose a novel method based on graph-theoretic framework. Crucial points to our approach are: a) it considers the sentence similarity globally, b) it enables us to capture the discriminative email features along with LDA and c) it combines the strengths of both supervised and unsupervised approaches.

4 Proposed Solution

Our proposed solution incorporates the discriminative features to identify topics and assign sentences to topics. We characterize each pair of sentences with: 1) Topic features (LSA, LDA), 2) Conversation features (distance between the two sentences in the fragment quotation graph, speaker, mention of names, time, subject of the email, "reply to" relation in email), and 3) Lexical features (tf*idf, Cue words). Note that we are using the output of the unsupervised methods (LSA and LDA) as features in the supervised binary classifier (described next).

Inspired by [5] we use a binary classifier learned from a small training set, to decide, given any two sentences, whether they should be in the same topic or not. In the next step we form an undirected graph $G = (V, E)$, where the nodes V represent the sentences of an email thread and the edge weights $w(u, v)$ denote the class (i.e., same topic) membership probability for the two sentences u and v. Note that the graph is completely connected. In this way, we are considering the similarity between the sentences globally.

Once we have the graph, similarly to [10], the problem of finding topics can be formulated as a graph partitioning problem. We aim to partition V into disjoint subsets $V_1, V_2 \dots V_m$, where the similarity among the vertices in a subset V_i is high and, across different subsets V_i, V_j is low. Here, we propose to use the normalized cut criterion which has been successfully applied in computer vision for image segmentation [13]. The normalized cut criterion is: $Ncut(A, B) = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(B,A)}{assoc(B,V)}$, where $assoc(A, V) = \Sigma_{u \in A, t \in V} w(u, t)$ is the total connection from nodes in partition A to all nodes in the graph and $assoc(B, V)$ is similarly defined. Solving this problem turns out to be NP-hard. Hence, we try to approximate the solution to optimize the normalized cut criterion following [13].

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[2] J. Boyd-Graber and D. Blei. Syntactic topic models. In *Neural Information Processing Systems*, 2008.

[3] G. Carenini, R. T. Ng, and X. Zhou. Summarizing email conversations with clue words. In *Proceedings of the 16th international conference on World Wide Web*, pages 91–100. ACM New York, NY, USA, 2007.

[4] H. Chen, S. R. K. Branavan, R. Barzilay, and D. R. Karger. Global models of document structure using latent permutations. In *NAACL'09*, pages 371–379, Morristown, NJ, USA, 2009. ACL.

[5] M. Elsner and E. Charniak. You talking to me? a corpus and algorithm for conversation disentanglement. In *Proceedings of ACL-08: HLT*, pages 834–842, Ohio, June 2008. ACL.

[6] M. Galley, K. Mckeown, E. F. Lussier, and H. Jing. Discourse segmentation of multi-party conversation. In *in 41st Annual Meeting of ACL*, pages 562–569, 2003.

[7] M. Georgescul, A. Clark, and S. Armstrong. A comparative study of mixture models for automatic topic segmentation of multiparty dialogues. In *ACL-08:HLT*, pages 925–930, Ohio, June 2008. ACL.

[8] T. L. Griffiths and M. Steyvers. Prediction and semantic association. In *Advances in Neural Information Processing Systems*. MIT Press, 2003.

[9] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum. Integrating topics and syntax. In *In Advances in Neural Information Processing Systems*, pages 537–544. MIT Press, 2005.

[10] I. Malioutov and R. Barzilay. Minimum cut model for spoken lecture segmentation. In *Proceedings of the ACL'06*, pages 25–32, Sydney, Australia, July 2006. ACL.

[11] J. O'Neill and D. Martin. Text chat in action. In *GROUP '03: Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work*, pages 40–49, New York, NY, USA, 2003. ACM.

[12] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on UAI*, pages 487–494, VA, USA, 2004. AUAI Press.

[13] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[14] M. Steyvers and T. Griffiths. *Latent Semantic Analysis: A Road to Meaning*, chapter Probabilistic topic models. Laurence Erlbaum, 2007.