Who Talks to Whom: Modeling Latent Structures in Dialogue Documents

Bailu Ding[1], **Jiang-Ming Yang**[2], **Chong Wang**[3], **Rui Cai**[2], **Zhiwei Li**[2], **Lei Zhang**[2]
[1]Fudan University, bailuding@gmail.com
[2]Microsoft Research Asia, {jmyang, ruicai, zli, leizhang}@microsoft.com
[3]Princeton University, chongw@cs.princeton.edu

1 Latent Structures of Dialogue Documents

Various forms of data that consist of sequential messages abound in social networks, such as citations, mail lists, chats, and forum discussions. A sequence of messages can be seen as a *dialogue*. We call the correlation among messages, namely the *'who-talks-to-whom'* relationship, the *dialogue structure*. Discovering dialogue structure is important for further studies on user behaviors.

In a dialogue, we call the first message the *root message*, and the member who posts the root message the *dialogue starter*. People interested in the root message 'talk' to the dialogue starter by posting new messages. Figure 1 (left) shows a dialogue taken from a discussion forum Slashdot(http://slashdot.org). Once we discover the tree-structured reply relationship in the dialogue, as shown in the upper-right part of Figure 1, we can construct the 'who-talks-to-whom' network as shown in the lower-right part of Figure 1. This network can be used in various applications, such as community identification, expert ranking and friend recommendation.

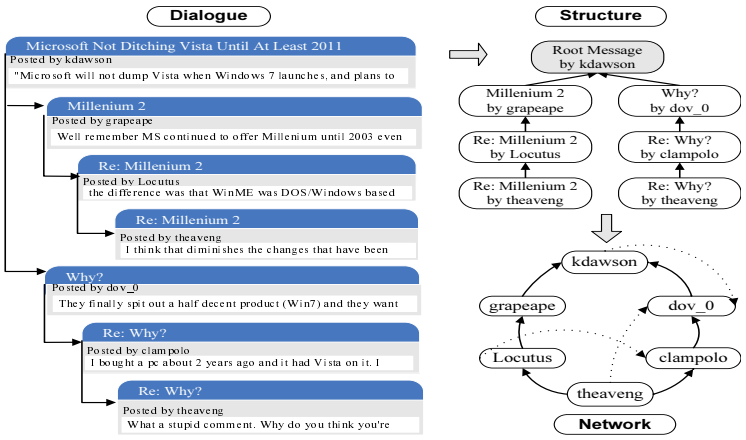


Figure 1: **Left**. A dialogue taken from the discussion forum Slashdot. **Right**. The upper part is the tree-structured reply relationship derived from the dialogue on the left. The lower part is the corresponding social network. Dotted arrows are edges derived from other dialogues.

With the exception of some limited scenarios, such as citations and mail lists, the explicit dialogue structure is generally unavailable. In cases like group chats, the functionality of replying to a specific member is absent. In cases like forums where partial information is provided, straightforward methods, such as leveraging online forum software, quoted text hints, users' name, common title words, are useful, but results are unsatisfactory. For example, people rarely bother using the 'quote' button provided by software(vBulletin, phpBB, etc.) generated forums. We randomly sampled 172765 messages from 20 popular forums, only 11.7% of them 'quote' some forms of partial information, like user name and partial message. Thus, the recovery of the latent dialogue structure is non-trivial.

We model this latent dialogue structure based on the following properties:

- Messages within a dialogue arise from **similar topics**, which suggests the opportunity to use topic models.

- The messages within a dialogue are **not exchangeable**. Therefore, most previous research work resting upon the premise of the exchangeability of the documents are inappropriate for modeling latent dialogue structure.
- A message may talk to **any previous messages** within a dialogue. Some recent papers on topic modeling incorporate the dynamics of documents [1, 2] , but they only capture the correlation between documents in one time slice and those in the following one.
- The correlation between messages is **sparse**. A member often replies to a single message, influenced by a small number of previous messages.

In this work, we present the *latent dialogue structure* (LDS) model, a family of topic models attempting to discover the latent dialogue structure. We design a special correlation matrix to model the relations of the messages within the dialogue. We further employ a Laplace prior to ensure sparsity. We show that this model outperforms other approaches in modeling latent dialogue structure.

2 Latent dialogue structure model

The LDS model treats the words of a message as arising from a set of latent topics. Messages in a dialogue share the same K topics, and each message uses a specific mixture proportion of topics.

We model the reply relationships with a special-structured correlation matrix. As previous messages influence new messages, correlated messages should have similar topic mixtures. We draw the topic mixture of a message from a linear combination of those of previous messages with Gaussian noise:

$$\boldsymbol{\eta}_m \sim N(\boldsymbol{t}_m^T \boldsymbol{\eta}_{<m}, \boldsymbol{\Sigma}_m) \tag{1}$$

where $\boldsymbol{\eta}_{<m} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_{m-1})$ are the topic mixtures of the previous messages, $\boldsymbol{\Sigma}_m$ is a covariance matrix of Gaussian noise, and $\boldsymbol{t}_m = (t_m^{(1)}, t_m^{(1)}, \ldots, t_m^{(m-1)})$ represents the coefficients of the linear combination of topic mixtures for message m. Each $t_m^{(i)} (1 \leq i \leq m-1)$ represents the correlation weight between message m and i.

To ensure sparsity, we draw $t_m^{(i)}$ from a Laplace prior with mean 0. The correlation matrix of a dialogue d is $\mathbf{T}^{(\mathbf{d})} = (\boldsymbol{t}_1^{(d)}, \boldsymbol{t}_2^{(d)}, \ldots, \boldsymbol{t}_M^{(d)})$. In practice, we use $\boldsymbol{\Sigma}_m = \delta^2 I$.

The LDS model assumes that a dialogue arises from the following generative process:

1. For dialogue d in corpus, $1 \leq d \leq D$:
2. For message m in the dialogue d, $1 \leq m \leq M_d$:
3. For component of $\boldsymbol{t}_m$, draw $t_m^{(i)} \sim$ Laplace$(0, b)$, where $1 \leq i < m$ and b is a scale prior.
4. Draw $\boldsymbol{\eta}_m \sim N(\boldsymbol{t}_m^T \boldsymbol{\eta}_{<m}, \boldsymbol{\Sigma}_m)$, where $\boldsymbol{\Sigma}_m$ is a Gaussian noise.
5. For each word $w_m^{(n)}$ in the message, where $1 \leq n \leq N_m$:
   (a) Draw $z_m^{(n)} | \boldsymbol{\eta}_m \sim$ Mult$(f(\boldsymbol{\eta}_m))$, where $f(\boldsymbol{\eta}_m) \propto \exp(\boldsymbol{\eta}_m)$.
   (b) Draw $w_m^{(n)} | z_m^{(n)}, \boldsymbol{\beta} \sim$ Mult$(\boldsymbol{\beta}_{z_m^{(n)}})$.

Here, D is the number of dialogues, M the number of messages, and M_d the number of messages in each dialogue d. The graphical model is depicted in Figure 2.

3 Inference and Estimation

The likelihood function of the whole data corpus is:

$$\prod_{d=1}^{D} \prod_{m=1}^{M_d} p(\boldsymbol{w}_m | b, \boldsymbol{t}_m^T \boldsymbol{\eta}_{<m}, \boldsymbol{\Sigma}_m, \boldsymbol{\beta}), \tag{2}$$

Since the dialogues are independent, we will only discuss the case with a single dialogue. Given a dialogue, the posterior distribution of the latent variables is

$$\begin{aligned} &p(\boldsymbol{\eta}_m, \boldsymbol{z}_m, \boldsymbol{t}_m | \boldsymbol{t}_m^T \boldsymbol{\eta}_{<m}, \boldsymbol{\Sigma}_m, \boldsymbol{\beta}, \boldsymbol{w}_m, b) \\ &= \frac{p(\boldsymbol{w}_m, \boldsymbol{\eta}_m, \boldsymbol{z}_m, \boldsymbol{t}_m | \boldsymbol{t}_m^T \boldsymbol{\eta}_{<m}, \boldsymbol{\Sigma}_m, \boldsymbol{\beta}, b)}{\int\int_{\boldsymbol{t}_m, \boldsymbol{\eta}_m} p(\boldsymbol{t}_m | 0, b) \sum_{\boldsymbol{z}_m} (p(\boldsymbol{\eta}_m | \boldsymbol{t}_m^T \boldsymbol{\eta}_{<m}, \boldsymbol{\Sigma}_m) \prod_{n=1}^{N_m} (p(z_m^{(n)} | \boldsymbol{\eta}_m) p(w_m^{(n)} | \boldsymbol{\beta}_{z_m^{(n)}})))}. \end{aligned} \tag{3}$$

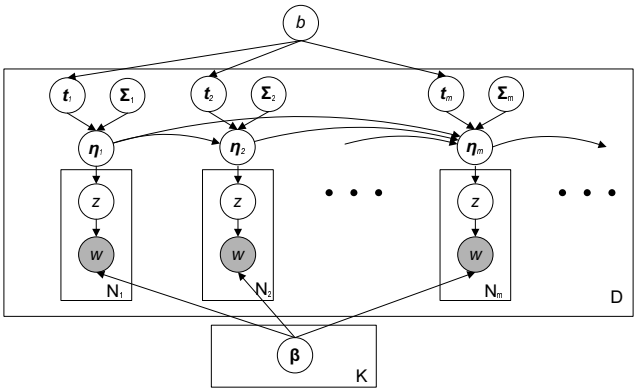


Figure 2: A graphical model representation of LDS model. D is the number of dialogues. M_d is the number of messages in each dialogue d. N_m represents the number of words in message m. K is the number of topics. The messages in all dialogues share the same topic parameter β.

Since the closed form is intractable, we utilize variational method to approximate this posterior distribution. We use the Jensen's inequality to approximate the lower bound of the log-likelihood:

$$\begin{aligned}
&\log(p(\boldsymbol{w}_m|b, \boldsymbol{t}_m^T\boldsymbol{\eta}_{<m}, \boldsymbol{\Sigma}_m, \boldsymbol{\beta})) \\
&\geq L(\boldsymbol{t}_m^T\boldsymbol{\eta}_{<m}, \boldsymbol{\eta}_m, \boldsymbol{\Sigma}_m, b, \boldsymbol{\beta}|\alpha_m, \gamma_m, \boldsymbol{\lambda}_m, \boldsymbol{V}_m, \boldsymbol{\Phi}_m) \\
&= E_q[\log(p(\boldsymbol{t}_m|0, b))] + E_q[\log(p(\boldsymbol{\eta}_m|\boldsymbol{t}_m^T\boldsymbol{\eta}_{<m}, \boldsymbol{\Sigma}_m)] \\
&+ \textstyle\sum_{n=1}^{N_m} E_q[\log(p(z_m^{(n)}|\boldsymbol{\eta}_m))] + \sum_{n=1}^{N_m} E_q[\log(p(w_m^{(n)}|\boldsymbol{\beta})] - H(q),
\end{aligned} \tag{4}$$

where the expectation is taken with respect to a variational distribution q. We fully factorize q as

$$\begin{aligned}
&q(\boldsymbol{t}_m, \boldsymbol{\eta}_m, \boldsymbol{z}_m|\boldsymbol{\alpha}_m, \boldsymbol{\gamma}_m, \boldsymbol{\lambda}_m, \boldsymbol{V}_m, \boldsymbol{\Phi}_m) \\
&= \textstyle\prod_{i=1}^{<m} q(t_m^{(i)}|\alpha_m^{(i)}, \gamma_m^{(i)}) \cdot \prod_{j=1}^{K} q(\boldsymbol{\eta}_m^{(j)}|\lambda_m^{(j)}, V_m^{(j,j)}) \cdot \prod_{k=1}^{N_m} q(z_m^{(k)}|\Phi_m^{(z_m^{(k)})}),
\end{aligned} \tag{5}$$

where $\{\boldsymbol{\alpha}_m, \boldsymbol{\gamma}_m\}$ are the variational parameters of Laplace distribution, $\{\boldsymbol{\lambda}_m, \boldsymbol{V}_m\}$ are the variational parameters of Gaussian distribution, and each $\boldsymbol{\Phi}_m$ specifies the variational distribution of the topic assignments $\boldsymbol{z}_m$.

The first and second term of equation (4) is

$$\begin{aligned}
&E_q[\log(p(\boldsymbol{t}_m|0, b)] = (1-m)\log(2b) \\
&- b^{-1}\textstyle\sum_{i=1}^{<m}(\frac{|\alpha_m^{(i)}|+\alpha_m^{(i)}}{2} + \frac{\gamma_m^{(i)}}{2} \cdot (exp(-\frac{|\alpha_m^{(i)}|}{\gamma_m^{(i)}}) + \exp(-\frac{\alpha_m^{(i)}}{\gamma_m^{(i)}}))).
\end{aligned} \tag{6}$$

$$\begin{aligned}
&E_q[\log(p(\boldsymbol{\eta}_m|\boldsymbol{t}_m^T\boldsymbol{\eta}_{<m}, \boldsymbol{\Sigma}_m))] \\
&= -(\boldsymbol{\lambda}_m - \boldsymbol{\alpha}_m^T\boldsymbol{\eta}_{<m})^T\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{\lambda}_m - \boldsymbol{\alpha}_m^T\boldsymbol{\eta}_{<m}) - \frac{1}{2}Tr(\boldsymbol{V}_m\boldsymbol{\Sigma}_m^{-1}) \\
&- \sum_{j=1}^{m-1}(\gamma_m^{(j)})^2(\boldsymbol{\eta}_j^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\eta}_j) - \frac{1}{2}\sum_{j=1}^{m-1}\sum_{s=1, s\neq j}^{m-1}\alpha_m^{(j)}\alpha_m^{(s)}(\boldsymbol{\eta}_j^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\eta}_s) - \frac{1}{2}\sum_{i=1}^{K}\log(2\pi\Sigma_m^{(i,i)}).
\end{aligned} \tag{7}$$

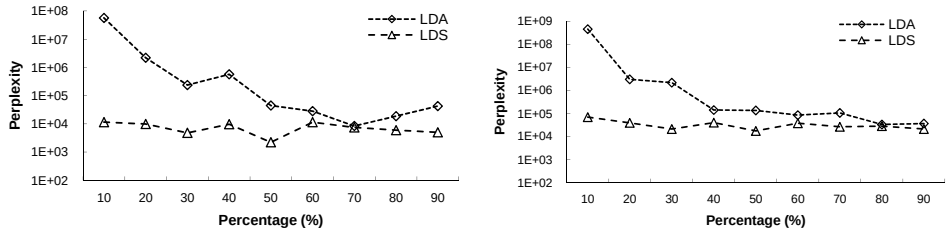
Note that the term $(\boldsymbol{\lambda}_m - \boldsymbol{\alpha}_m^T\boldsymbol{\eta}_{<m})^T\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{\lambda}_m - \boldsymbol{\alpha}_m^T\boldsymbol{\eta}_{<m})$ is just the square of Mahalanobis distance, which penalizes the deviation between the approximation of the topic distribution of message m and the linear combination of those of previous messages. Other terms of equation (4) are easy to get.

4 Experiments

We ran a 15-topic LDS on two forums, Apple Discussion (http://discussions.apple.com) and Slashdot. These two forums provide explicit reply relationships that can be used as the ground truth for evaluation. Our dataset consists of 100 dialogues from Apple with 10366 messages, and 100 dialogues from Slashdot with 19005 messages.

Table 1: Results of modeling latent dialogue structure in Apple Discussion and Slashdot with TF-IDF, LDA, LDS, LDA+TF-IDF and LDS+TF-IDF averaged over 100 threads for each discussion board. Each thread has over 100 messages on average. Top 5 candidates are considered to reach a reasonable recall. The top two results are shown in bold.

Apple	**top 1**	**top 2**	**top 3**	**top 4**	**top 5**
TF-IDF	12.78%	18.86%	23.45%	27.58%	31.61%
LDA	8.24%	13.03%	16.65%	19.67%	22.33%
LDS	**27.69%**	**31.96%**	**35.66%**	**38.85%**	**41.36%**
LDA+TF-IDF	12.34%	18.16%	22.48%	26.20%	29.89%
LDS+TF-IDF	**28.03%**	**34.31%**	**38.19%**	**41.30%**	**44.22%**
Slashdot	**top 1**	**top 2**	**top 3**	**top 4**	**top 5**
TF-IDF	**25.02%**	**35.30%**	**41.88%**	**47.12%**	**51.01%**
LDA	9.07%	14.83%	19.27%	23.18%	26.46%
LDS	17.97%	22.97%	27.93%	31.98%	35.42%
LDA+TF-IDF	24.84%	34.37%	40.56%	45.15%	48.60%
LDS+TF-IDF	**29.78%**	**39.67%**	**45.48%**	**50.25%**	**53.71%**



(a) Apple Discussion (b) Slashdot

Figure 3: Predicting ability of LDS and LDA in Apple Discussion and Slashdot. We tested on different proportions of seen data from 10 to 90 percent to predict the next 10 percent messages.

4.1 Revealing dialogue structure

In this experiment, we first ran variational EM until the relative change in the likelihood bound is less than 10^{-5}. For each message pair i and j, we computed the cosine distance between $\boldsymbol{t}_i$ and $\boldsymbol{t}_j$. As $\boldsymbol{t}_i$ and $\boldsymbol{t}_j$ are of different lengths, we use their average and set extra components to 0. We ranked the candidate reply messages w.r.t their cosine distance. In practice, we use $\boldsymbol{\alpha}$ to approximate $\boldsymbol{t}$.

We also ran TF-IDF and LDA on our datasets. We used the same number of topics and convergence condition in LDA as we did in LDS. Then we computed the cosine distance of topics distribution vector in LDA and the cosine distance of words distribution vector in TF-IDF. We further tested combinations of LDA+TF-IDF and LDS+TF-IDF to produce more robust models. We assigned a ratio λ and $(1-\lambda)$ for the combination of two models. We chose the value of λ empirically based on a small subset of data; 0.1 is used for LDA+TF-IDF and $\lambda = 0.5$ for LDS+TF-IDF. We chose top ranked messages as candidates, and computed the recall w.r.t the ground-truth message.

The results are shown in Table 1. When messages are short, as in Apple Discussion, LDS alone produces reasonable results. When messages are long and more informative, as in Slashdot, TF-IDF performs better, but its performance can be greatly improved with the information of $\boldsymbol{t}$. The combined model of LDS and TF-IDF is more robust in corpora of different styles.

4.2 Predicting unseen messages

We evaluated the prediction abilities of LDS and LDA. We ran variational inference until the change in the probability bound of equation (4) is less than 10^{-5}. Average perplexity per word is used as the evaluation criterion. The lower the perplexity, the better the performance. As figure 3 shows, LDS has a more powerful predictive ability than LDA.

References

[1] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the International Conference on Machine Learning*, pages 113–120, 2006.

[2] C. Wang, D. M. Blei, and D. Heckerman. Continuous time dynamic topic models. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2008.