# Multi-Document Summarization using Sentence-based Topic Models

**Dingding Wang** [1] **Shenghuo Zhu** [2] **Tao Li** [1] **Yihong Gong** [2]
1. School of Computer Science, Florida International University, Miami, FL, 33199
2. NEC Laboratories America, Cupertino, CA 95014, USA.
{dwang003,taoli}@cs.fiu.edu {zsh,ygong}@sv.nec-labs.com

## Abstract

Most of the existing multi-document summarization methods decompose the documents into sentences and work directly in the sentence space using a term-sentence matrix. However, the knowledge on the document side, i.e. the topics embedded in the documents, can help the context understanding and guide the sentence selection in the summarization procedure. In this paper, we propose a new Bayesian sentence-based topic model for summarization by making use of both the term-document and term-sentence associations. An efficient variational Bayesian algorithm is derived for model parameter estimation. Experimental results on benchmark data sets show the effectiveness of the proposed model for the multi-document summarization task.

## 1 Introduction

With the continuing growth of online text resources, document summarization has found wide-ranging applications in information retrieval and web search. Many multi-document summarization methods have been developed to extract the most important sentences from the documents. These methods usually represent the documents as term-sentence matrices (where each row represents a sentence and each column represents a term) or graphs (where each node is a sentence and each edge represents the pairwise relationship among corresponding sentences), and ranks the sentences according to their scores calculated by a set of predefined features, such as term frequency-inverse sentence frequency (TF-ISF) (Radev et al., 2004; Lin and Hovy, 2002), sentence or term position (Yih et al., 2007), and number of keywords (Yih et al., 2007). Typical existing summarization methods include centroid-based methods (e.g., MEAD (Radev et al., 2004)), graph-ranking based methods (e.g., LexPageRank (Erkan and Radev, 2004)), non-negative matrix factorization (NMF) based methods (e.g., (Lee and Seung, 2001)), Conditional random field (CRF) based summarization (Shen et al., 2007), and LSA based methods (Gong and Liu, 2001).

There are two limitations with most of the existing multi-document summarization methods: (1) They work directly in the sentence space and many methods treat the sentences as independent of each other. Although few work tries to analyze the context or sequence information of the sentences, the document side knowledge, i.e. the topics embedded in the documents are ignored. (2) Another limitation is that the sentence scores calculated from existing methods usually do not have very clear and rigorous probabilistic interpretations. Many if not all of the sentence scores are computed using various heuristics as few research efforts have been reported on using generative models for document summarization.

In this paper, to address the above issues, we propose a new Bayesian sentence-based topic model for multi-document summarization by making use of both the term-document and term-sentence associations. Our proposal explicitly models the probability distributions of selecting sentences given topics and provides a principled way for the summarization task. An efficient variational Bayesian algorithm is derived for estimating model parameters.

## 2 Bayesian Sentence-based Topic Models (BSTM)

### 2.1 Model Formulation

The entire document set is denoted by $\mathcal{D}$. For each document $d \in \mathcal{D}$, we consider its unigram language model,

*Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 297–300,
Suntec, Singapore, 4 August 2009. ©2009 ACL and AFNLP

$$p(W_1^n|\boldsymbol{\theta}_d) = \prod_{i=1}^{n} p(W_i|\boldsymbol{\theta}_d),$$

where $\boldsymbol{\theta}_d$ denotes the model parameter for document $d$, $W_1^n$ denotes the sequence of words $\{W_i \in \mathcal{W}\}_{i=1}^n$, i.e. the content of the document. $\mathcal{W}$ is the vocabulary. As topic models, we further assume the unigram model as a mixture of several topic unigram models,

$$p(W_i|\boldsymbol{\theta}_d) = \sum_{T_i \in \mathcal{T}} p(W_i|T_i)p(T_i|\boldsymbol{\theta}_d),$$

where $\mathcal{T}$ is the set of topics. Here, we assume that given a topic, generating words is independent from the document, i.e.

$$p(W_i|T_i, \boldsymbol{\theta}_d) = p(W_i|T_i).$$

Instead of freely choosing topic unigram models, we further assume that topic unigram models are mixtures of some existing *base unigram models*, i.e.

$$p(W_i|T_i) = \sum_{s \in \mathcal{S}} p(W_i|S_i = s)p(S_i = s|T_i),$$

where $\mathcal{S}$ is the set of base unigram models. Here, we use sentence language models as the base models. One benefit of this assumption is that each topic is represented by meaningful sentences, instead of directly by keywords. Thus we have

$$p(W_i|\boldsymbol{\theta}_d) = \sum_{t \in \mathcal{T}} \sum_{s \in \mathcal{S}} p(W_i|S_i = s)p(S_i = s|T_i = t)p(T_i = t|\boldsymbol{\theta}_d).$$

Here we use parameter $U_{st}$ for the probability of choosing base model $s$ given topic $t$, $p(S_i = s|T_i = t) = U_{st}$, where $\sum_s U_{st} = 1$. We use parameters $\{\boldsymbol{\theta}_d\}$ for the probability of choosing topic $t$ given document $d$, where $\sum_t \Theta_{dt} = 1$. We assume that the parameters of base models, $\{B_{ws}\}$, are given, i.e. $p(W_i = w|S_i = s) = B_{ws}$, where $\sum_w B_{ws} = 1$. Usually, we obtain $B_{ws}$ by empirical distribution words of sentence $s$.

### 2.2 Parameter Estimation

For summarization task, we concern how to describe each topic with the given sentences. This can be answered by the parameter of choosing base model $s$ given topic $t$, $U_{st}$. Comparing to parameter $U_{st}$, we concern less about the topic distribution of each document, i.e. $\Theta_{dt}$. Thus we choose Bayesian framework to estimate $U_{st}$ by marginalizing $\Theta_{dt}$. To do so, we assume a Dirichlet prior for $\boldsymbol{\Theta}_{d\cdot} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, where vector $\boldsymbol{\alpha}$ is a hyperparameter. Thus the likelihood is

$$\begin{aligned} f(\mathbf{U};\mathbf{Y}) &= \prod_d \int \prod_i p(Y_{id}|\boldsymbol{\theta}_d)\pi(\boldsymbol{\theta}_d|\boldsymbol{\alpha})d\boldsymbol{\theta}_d \\ &= B(\boldsymbol{\alpha})^{-D} \int \prod_{id} [\mathbf{B}\mathbf{U}\boldsymbol{\Theta}^\top]_{id}^{Y_{id}} \times \prod_{dk} \Theta_{dk}^{\alpha_k - 1} d\,\boldsymbol{\Theta}. \end{aligned} \tag{1}$$

As Eq. (1) is intractable, LDA (Blei et al., 2001) applies variational Bayesian, which is to maximize a variational bound of the integrated likelihood. Here we write the variational bound.

**Definition 1** *The variational bound is*

$$\widetilde{f}(\mathbf{U},\mathbf{V};\mathbf{Y}) = \prod_d \frac{B(\boldsymbol{\alpha} + \boldsymbol{\gamma}_{d,\cdot})}{B(\boldsymbol{\alpha})} \prod_{vkwd} \left(\frac{B_{wv}U_{vk}}{\phi_{vk;wd}}\right)^{Y_{wd}\phi_{vk;wd}} \tag{2}$$

*where the domain of* $\mathbf{V}$ *is*$\mathcal{V} = \{\mathbf{V} \in \mathbb{R}_+^{D\times K} : \sum_k V_{dk} = 1\}$, $\phi_{vk;wd} = B_{wv}U_{vk}V_{dk}/[\mathbf{B}\mathbf{U}\mathbf{V}^\top]_{wd}$, $\gamma_{dk} = \sum_{wv} Y_{wd}\phi_{vk;wd}$.

We have the following proposition.

**Proposition 1** $f(\mathbf{U};\mathbf{Y}) \geq \sup_{\mathbf{V}\in\mathcal{V}} \widetilde{f}(\mathbf{U},\mathbf{V};\mathbf{Y})$.

Actually the optimum of this variational bound is the same as that obtained variational Bayesian approach. Due to the space limit, the proof of the proposition is omitted.

## 3 The Iterative Algorithm

The LDA algorithm (Blei et al., 2001) employed the variational Bayesian paradigm, which estimates the optimal variation bound for each $\mathbf{U}$. The algorithm requires an internal Expectation-Maximization (EM) procedure to find the optimal variational bound. The nested EM slows down the optimization procedure. To avoid the internal EM loop, we can directly optimize the variational bound to obtain the update rules.

### 3.1 Algorithm Derivation

First, we define the concept of Dirichlet adjustment, which is used in the algorithm for variational update rules involving Dirichlet distribution. Then, we define some notations for the update rules.

**Definition 2** *We call vector* $\mathbf{y}$ *of size* $K$ *is the* Dirichlet adjustment *of vector* $\mathbf{x}$ *of size* $K$ *with respect to Dirichlet distribution* $D_K(\boldsymbol{\alpha})$ *if*

$$y_k = \exp(\Psi(\alpha_k + x_k) - \Psi(\sum_l (\alpha_l + x_l))),$$

*where* $\Psi(\cdot)$ *is digamma function. We denote it by* $\mathbf{y} = \mathcal{P}_D(\mathbf{x};\boldsymbol{\alpha})$.

We denote element-wise product of matrix $\mathbf{X}$ and matrix $\mathbf{Y}$ by $\mathbf{X} \circ \mathbf{Y}$, element-wise division by $\frac{\mathbf{X}}{\mathbf{Y}}$, obtaining $\mathbf{Y}$ via normalizing of each column of $\mathbf{X}$ as $\mathbf{Y} \overset{1}{\leftarrow} \mathbf{X}$, and obtaining $\mathbf{Y}$ via Dirichlet adjustment $\mathcal{P}_D(\cdot;\boldsymbol{\alpha})$ and normalization of each row of $\mathbf{X}$ as $\overset{\mathcal{P}_D(\cdot;\boldsymbol{\alpha}),2}{\longleftarrow}$, i.e., $\mathbf{z} = \mathcal{P}_D((\mathbf{X}_{d,\cdot})^\top;\boldsymbol{\alpha})$ and $Y_{d,k} = z_k/\sum_k z_k$. The following is the update rules for LDA:

$$\mathbf{U} \overset{1}{\leftarrow} \mathbf{B}^\top \left[\frac{\mathbf{Y}}{\mathbf{B}\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^\top}\right] \widetilde{\mathbf{V}} \circ \widetilde{\mathbf{U}} \tag{3}$$

$$\mathbf{V} \overset{\mathcal{P}_D(\cdot;\boldsymbol{\alpha}),2}{\longleftarrow} \left[\frac{\mathbf{Y}}{\mathbf{B}\mathbf{U}\widetilde{\mathbf{V}}^\top}\right]^\top (\mathbf{B}\mathbf{U}) \circ \widetilde{\mathbf{V}} \tag{4}$$

**Algorithm 1** Iterative Algorithm

```
Input:   Y   : term-document matrix
         B   : term-sentence matrix
         K   : the number of latent topics
Output:  U   : sentence-topic matrix
         V   : auxiliary document-topic matrix

Randomly initialize U and V, and normalize them
repeat
    Update U using Eq. (3);
    Update V using Eq. (4);
    Compute $\widetilde{f}$ using Eq. (2);
until $\widetilde{f}$ converges.
```

### 3.2 Algorithm Procedure

The detail procedure is listed as Algorithm 1. ¿From the sentence-topic matrix $\mathbf{U}$, we include the sentence with the highest probability in each topic into the summary.

## 4 Relations with Other Models

In this section, we discuss the connections and differences of our BSTM model with two related models.

Recently, a new language model, factorization with sentence bases (FGB) (Wang et al., 2008) is proposed for document clustering and summarization by making use of both term-document matrix $\mathbf{Y}$ and term-sentence matrix $\mathbf{B}$. The FGB model computes two matrices $\mathbf{U}$ and $\mathbf{V}$ by optimizing

$$\mathbf{U}, \mathbf{V} = \arg\min_{\mathbf{U},\mathbf{V}} \ell(\mathbf{U}, \mathbf{V}),$$

where

$$\ell(\mathbf{U}, \mathbf{V}) = \mathsf{KL}\left(\mathbf{Y} \| \mathbf{B}\mathbf{U}\mathbf{V}^\top\right) - \ln \Pr(\mathbf{U}, \mathbf{V}).$$

Here, Kullback-Leibler divergence is used to measure the difference between the distributions of $\mathbf{Y}$ and the estimated $\mathbf{B}\mathbf{U}\mathbf{V}^\top$. Our BSTM is similar to the FGB summarization since they are all based on sentence-based topic model. The difference is that the document-topic allocation $\mathbf{V}$ is marginalized out in BSTM. The marginalization increases the stability of the estimation of the sentence-topic parameters. Actually, from the algorithm we can see that the difference lies in the Dirichlet adjustment. Experimental results show that our BSTM achieves better summarization results than FGB model.

Our BSTM model is also related to 3-factor non-negative matrix factorization (NMF) model (Ding et al., 2006) where the problem is to solve $\mathbf{U}$ and $\mathbf{V}$ by minimizing

$$\ell_F(\mathbf{U}, \mathbf{V}) = \|\mathbf{Y} - \mathbf{B}\mathbf{U}\mathbf{V}^\top\|_F^2. \tag{5}$$

Both BSTM and NMF models are used for solving $\mathbf{U}$ and $\mathbf{V}$ and have similar multiplicative update rules. Note that if the matrix $\mathbf{B}$ is the identity matrix, Eq. (5) leads to the derivation of the NMF algorithm with Frobenius norm in (Lee and Seung, 2001). However, our BSTM model is a generative probabilistic model and makes use of Dirichlet adjustment. The results obtained in our model have clear and rigorous probabilistic interpretations that the NMF model lacks. In addition, by marginalizing out $\mathbf{V}$, our BSTM model leads to better summarization results.

## 5 Experimental Results

### 5.1 Data Set

To evaluate the summarization results empirically, we use the DUC2002 and DUC2004 data sets, both of which are open benchmark data sets from Document Understanding Conference (DUC) for generic automatic summarization evaluation. Table 1 gives a brief description of the data sets.

| | DUC2002 | DUC2004 |
|---|---|---|
| number of document collections | 59 | 50 |
| number of documents in each collection | ∼10 | 10 |
| data source | TREC | TDT |
| summary length | 200 words | 665bytes |

Table 1: Description of the data sets for multi-document summarization

| Systems | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-SU |
|---|---|---|---|---|
| DUC Best | **0.49869** | **0.25229** | **0.46803** | **0.28406** |
| Random | 0.38475 | 0.11692 | 0.37218 | 0.18057 |
| Centroid | 0.45379 | 0.19181 | 0.43237 | 0.23629 |
| LexPageRank | 0.47963 | 0.22949 | 0.44332 | 0.26198 |
| LSA | 0.43078 | 0.15022 | 0.40507 | 0.20226 |
| NMF | 0.44587 | 0.16280 | 0.41513 | 0.21687 |
| KM | 0.43156 | 0.15135 | 0.40376 | 0.20144 |
| FGB | 0.48507 | 0.24103 | 0.45080 | 0.26860 |
| BSTM | **0.48812** | **0.24571** | **0.45516** | **0.27018** |

Table 2: Overall performance comparison on DUC2002 data using ROUGE evaluation methods.

| Systems | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-SU |
|---|---|---|---|---|
| DUC Best | **0.38224** | **0.09216** | **0.38687** | **0.13233** |
| Random | 0.31865 | 0.06377 | 0.34521 | 0.11779 |
| Centroid | 0.36728 | 0.07379 | 0.36182 | 0.12511 |
| LexPageRank | 0.37842 | 0.08572 | 0.37531 | 0.13097 |
| LSA | 0.34145 | 0.06538 | 0.34973 | 0.11946 |
| NMF | 0.36747 | 0.07261 | 0.36749 | 0.12918 |
| KM | 0.34872 | 0.06937 | 0.35882 | 0.12115 |
| FGB | 0.38724 | 0.08115 | 0.38423 | 0.12957 |
| BSTM | **0.39065** | **0.09010** | **0.38799** | **0.13218** |

Table 3: Overall performance comparison on DUC2004 data using ROUGE evaluation methods.

### 5.2 Implemented Systems

We implement the following most widely used document summarization methods as the baseline systems to compare with our proposed BSTM method. (1) Random: The method selects sentences randomly for each document collection.

(2) Centroid: The method applies MEAD algorithm (Radev et al., 2004) to extract sentences according to the following three parameters: centroid value, positional value, and first-sentence overlap. (3) LexPageRank: The method first constructs a sentence connectivity graph based on cosine similarity and then selects important sentences based on the concept of eigenvector centrality (Erkan and Radev, 2004). (4) LSA: The method performs latent semantic analysis on terms by sentences matrix to select sentences having the greatest combined weights across all important topics (Gong and Liu, 2001). (5) NMF: The method performs non-negative matrix factorization (NMF) on terms by sentences matrix and then ranks the sentences by their weighted scores (Lee and Seung, 2001). (6) KM: The method performs K-means algorithm on terms by sentences matrix to cluster the sentences and then chooses the centroids for each sentence cluster. (7) FGB: The FGB method is proposed in (Wang et al., 2008).

### 5.3 Evaluation Measures

We use ROUGE toolkit (version 1.5.5) to measure the summarization performance, which is widely applied by DUC for performance evaluation. It measures the quality of a summary by counting the unit overlaps between the candidate summary and a set of reference summaries. The full explanation of the evaluation toolkit can be found in (Lin and E.Hovy, 2003). In general, the higher the ROUGE scores, the better summarization performance.

### 5.4 Result Analysis

Table 2 and Table 3 show the comparison results between BSTM and other implemented systems. From the results, we have the follow observations: (1) Random has the worst performance. The results of LSA, KM, and NMF are similar and they are slightly better than those of Random. Note that LSA and NMF provide continuous solutions to the same K-means clustering problem while LSA relaxes the non-negativity of the cluster indicator of K-means and NMF relaxes the orthogonality of the cluster indicator (Ding and He, 2004; Ding et al., 2005). Hence all these three summarization methods perform clustering-based summarization: they first generate sentence clusters and then select representative sentences from each sentence cluster. (2) The Centroid system outperforms clustering-based summarization methods in most cases. This is mainly because the Centroid based algorithm takes into account positional value and first-sentence overlap which are not used in clustering-based summarization. (3) LexPageRank outperforms Centroid. This is due to the fact that LexPageRank ranks the sentence using eigenvector centrality which implicitly accounts for information subsumption among all sentences (Erkan and Radev, 2004). (4) FGB performs better than LexPageRank. Note that FGB model makes use of both term-document and term-sentence matrices. Our BSTM model outperforms FGB since the document-topic allocation is marginalized out in BSTM and the marginalization increases the stability of the estimation of the sentence-topic parameters. (5) Our BSTM method outperforms all other implemented systems and its performance is close to the results of the best team in the DUC competition. Note that the good performance of the best team in DUC benefits from their preprocessing on the data using deep natural language analysis which is not applied in our implemented systems.

The experimental results provide strong evidence that our BSTM is a viable method for document summarization.

**Acknowledgement:** The work is partially supported by NSF grants IIS-0546280, DMS-0844513 and CCF-0830659.

## References

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems 14*.

C. Ding and X. He. K-means clustering and principal component analysis. In *Prodeedings of ICML 2004*.

Chris Ding, Xiaofeng He, and Horst Simon. 2005. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of Siam Data Mining*.

Chris Ding, Tao Li, Wei Peng, and Haesun Park. 2006. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of SIGKDD 2006*.

G. Erkan and D. Radev. 2004. Lexpagerank: Prestige in multi-document text summarization. In *Proceedings of EMNLP 2004*.

Y. Gong and X. Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of SIGIR*.

Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13*.

C-Y. Lin and E.Hovy. Automatic evaluation of summaries using n-gram cooccurrence statistics. In *Proceedings of NLT-NAACL 2003*.

C-Y. Lin and E. Hovy. 2002. From single to multi-document summarization: A prototype system and its evaluation. In *Proceedings of ACL 2002*.

I. Mani. 2001. *Automatic summarization*. John Benjamins Publishing Company.

D. Radev, H. Jing, M. Stys, and D. Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing and Management*, pages 919–938.

B. Ricardo and R. Berthier. 1999. *Modern information retrieval*. ACM Press.

D. Shen, J-T. Sun, H. Li, Q. Yang, and Z. Chen. 2007. Document summarization using conditional random fields. In *Proceedings of IJCAI 2007*.

Dingding Wang, Shenghuo Zhu, Tao Li, Yun Chi, and Yihong Gong. 2008. Integrating clustering and multi-document summarization to improve document understanding. In *Proceedings of CIKM 2008*.

W-T. Yih, J. Goodman, L. Vanderwende, and H. Suzuki. 2007. Multi-document summarization by maximizing informative content-words. In *Proceedings of IJCAI 2007*.