LATENT DIRICHLET LEARNING FOR DOCUMENT SUMMARIZATION

Ying-Lang Chang and Jen-Tzung Chien

Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan 70101, ROC
{ylchang,chien}@chien.csie.ncku.edu.tw

ABSTRACT

Automatic summarization is developed to extract the representative contents or sentences from a large corpus of documents. This paper presents a new hierarchical representation of words, sentences and documents in a corpus, and infers the Dirichlet distributions for latent topics and latent themes in word level and sentence level, respectively. The sentence-based latent Dirichlet allocation (SLDA) is accordingly established for document summarization. Different from the vector space summarization, SLDA is built to fit the fine structure of text documents, and is specifically designed for sentence selection. SLDA acts as a sentence mixture model with a mixture of Dirichlet themes, which are used to generate the latent topics in observed words. The theme model is inherent to distinguish sentences in a summarization system. In the experiments, the proposed SLDA outperforms other methods for document summarization in terms of precision, recall and F-measure.

Index Terms— latent Dirichlet allocation, language model, sentence extraction, document summarization

1. INTRODUCTION

As the internet grows prosperously, the amount of multimedia documents is excessively increased. Summarization can help readers quickly capture the theme and concept of the whole documents, and effectively save reading time. However, abstracting or extracting summary from a huge corpus needs a lot of manpower. How to develop an automatic summarization system becomes an important research topic. In general, the abstraction is a rewrite summary for a full document, while the extraction is to select the representative sentences into the summary so as to condense the original text data. The abstraction is too difficult and arduous, so mostly we focus on the extraction method. In [2], a model-based relevance measure between sentence and document was proposed. Here, we present a model-based approach to extract informative sentences for document summarization.

Automatic summarization is usually performed in two ways. One is concept-based summarization, and the other is query-based summarization. The former case directly extracts the sentences, which are related to the theme or gist of the original document while the latter case selects the sentences according to user queries, so as to fit the interests of users. Also, we perform the multi-document summarization where the sentences are selected across different documents. This case differs from the single document summarization because the concept and diversity in all documents should be modeled. In [10], a centroid-based summarization from multiple documents was addressed. The document clustering was executed to find centroid terms in each cluster where the relevance measure between sentence and centroid was calculated. Automatic summarization is not only applied for text documents but also for web pages and spoken documents. In [5][7], the speech-to-speech and speech-to-text summarization was developed. A two-stage summarization method consisting of sentence extraction and sentence compaction was presented. In [6], the vector space model (VSM) and latent semantic analysis (LSA) were applied to calculate the similarity between sentence and document. To deal with the problems of polysemy and synonym, the latent Dirichlet allocation (LDA) [1] was presented. However, the original model was not suitable for document summarization. LDA was extended to a latent Dirichlet co-clustering model for characterizing the hierarchy of a text corpus [11]. The Gibbs sampling was employed in parameter inference but with slow convergence.

In this study, we develop the *latent Dirichlet learning* approach to concept-based summarization from multiple text documents. This approach can be extended to a query-based summarization from text and speech documents. The framework of sentence-based latent Dirichlet allocation (SLDA) is established. The hierarchy of text corpus is compactly represented by the associated *sentence-based language models* for the application of document summarization. The Bayesian variational inference scheme is adopted to infer the variational model distributions as well as to estimate the SLDA parameters. A set of experiments are reported by using evaluation tools and measures. The performance of proposed SLDA compared to VSM and LDA summarization is illustrated. In what follows, we survey the related models of VSM and LDA. In section 3, the SLDA summarization and the parameter inference in SLDA are described. The experiments on document summarization are reported in section 5 and the conclusion is given in section 6.

2. RELATED MODELS

2.1 Vector space model

Sentence selection is a widely-accepted approach to make summary from a large corpus. The whole document is decomposed into individual sentences. Using vector space summarization [6], the sentence is sequentially selected and put into summary according to a relevance score between each sentence and the whole document. The weighted term-frequencies of individual words in sentence and document are used to form the vectors. The inner product between sentence and document vectors is computed as the relevance measure. In

the sequential selection, the document vector is continuously recomputed by eliminating the sentence, which has been selected. The updated document vector is employed to calculate the relevance score in subsequent sentence selection. In [6], VSM summarization was upgraded to LSA summarization where the sentence and document were projected to a low-dimensional latent semantic space. The sentences with high index values were selected.

2.2 Latent Dirichlet allocation

More attractively, Blei et al. [1] presented the latent Dirichlet allocation (LDA) for document representation. LDA was an extension of probabilistic LSA (PSLA) [3][8]. LDA performed better than PLSA due to its generalization to unseen documents. Figure 1 displays the graphical model of LDA document model. There are N words in a document, V vocabulary words, K latent topics and M documents in the corpus. Each word w of a document d is associated with a hidden variable z which represents the latent topic. Variable z is sampled from a multinomial distribution with parameter θ indicating the probability of latent topic. The prior density of multinomial parameter θ is given by a Dirichlet distribution with hyperparameter α . The $K \times V$ parameter matrix $\beta = \{\beta_{kw}\}$ denotes the topic language model. LDA outperformed PLSA and other latent topic models in evaluation of document model [1], and was applied to build LDA language model for speech recognition [4]. In this study, LDA is applied for document summarization.

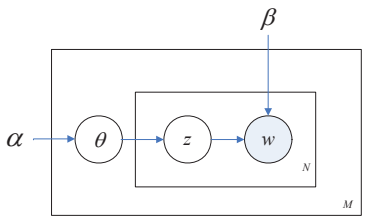

Figure 1: Graphical model for LDA.

The parameters of LDA $\{\alpha, \beta\}$ are estimated by maximizing a marginal likelihood $p(\mathbf{w} \mid \alpha, \beta)$ from a set of text documents $\mathbf{w} = \{w_{dn}\}$

$$\prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left[\prod_{n=1}^{N} \sum_{z_{dn}} p(w_{dn} \mid z_{dn}, \beta) p(z_{dn} \mid \theta_d) \right] d\theta_d \qquad (1)$$

where the marginalization is operated over latent topic $\mathbf{z} = \{z_{dn}\}$ and Dirichlet parameter θ_d . However, direct optimization of (1) is intractable. The variational inference is feasible to estimate LDA parameters by a *lower bound* of (1) as a surrogate for optimization. Considering the factored variational inference where $\mathbf{z}$ and θ are conditionally independent, a variational model $q(\theta, \mathbf{z} \mid \gamma, \phi)$ is formed to approximate the true posterior probability $p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta)$. By maximizing the lower bound, the variational parameters $\{\gamma, \phi\}$ and the LDA parameters $\{\alpha, \beta\}$ are estimated by

$$\phi_{nk} \propto \beta_{kw_n} \exp\{\Psi(\gamma_k) - \Psi(\sum_{j=1}^{K} \gamma_j)\} \qquad (2)$$

$$\gamma_k = \alpha_k + \sum_{n=1}^{N} \phi_{nk} \qquad (3)$$

$$\beta_{kw_n} \propto \sum_{d=1}^{M} \sum_{n=1}^{N} \phi_{dnk} w_{dn} \qquad (4)$$

$$\alpha^{t+1} = \alpha^{t} - H_{\mathrm{LDA}}(\alpha^{t})^{-1} g_{\mathrm{LDA}}(\alpha^{t}) \qquad (5)$$

where $\Psi(\cdot)$ is the first derivative of log Gamma function, t is the iteration index in decent algorithm, $H_{\mathrm{LDA}}(\cdot)$ and $g_{\mathrm{LDA}}(\cdot)$ denote the Hessian matrix and gradient vector of the lower bound with respect to α , respectively. The estimated variational model $q(\theta, \mathbf{z} \mid \phi, \gamma)$ approximates $p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta)$ with the smallest Kullback-Leibler (KL) divergence.

3. LATENT DIRICHLET SUMMARIZATION

3.1 Summarization by VSM and LDA

Summarization using VSM is performed by calculating inner product of sentence and document vectors. The rank list of sentences is obtained as the result of summary. Usually, VSM method is sensitive to the appearance of synonyms and co-occurrence words. In addition, LDA is a model-based approach where the text data are modeled in word and document levels. Sentence-based modeling is not considered. To implement the document summarization, LDA is not only performed in the whole document but also in individual sentences. Each sentence is viewed as a document in the implementation. As a result, we calculate LDA parameters of whole document and individual sentences. The rank list of sentences is generated by measuring KL divergence between document language model and sentence language model. However, neither VSM nor LDA tackles the sentence level modeling. The estimated models are not suitable for sentence representation. The hierarchy in words, sentences and documents is not sufficiently characterized, so the performance of sentence selection is limited. We are motivated to present the SLDA algorithm for document summarization.

3.2 Sentence-based latent Dirichlet allocation

SLDA is a hierarchical model with graphical representation shown in Figure 2. In this model, there are S sentences in a document. Each document d is modeled by a mixture of L latent themes. Each sentence s is associated with a latent theme η , which is a multinomial distribution with parameter θ . Each word w in a sentence s is sampled by a latent topic z, which is associated with a theme variable η in sentence level. SLDA process is described as follows

1. For each document d:
   - Choose the combination of themes by $\theta \sim \mathrm{Dir}(\alpha)$.
2. For each of S sentences s:
   - Choose a theme $\eta_{ds} \sim \mathrm{Multinomial}(\theta)$.
3. For each of N words w_{dsn} :
   - Choose a topic $z_{dsn} \mid \eta_{ds} \sim \mathrm{Multinomial}(\mu)$.
   - Choose a word $w_{dsn} \sim \mathrm{Multinomial}(\beta)$.

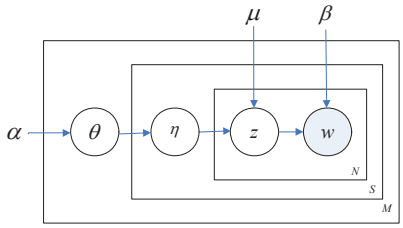

Figure 2: Graphical model for SLDA.

In contrast to LDA in (1), the marginal likelihood $p(\mathbf{w}\,|\,\alpha,\beta,\mu)$ of SLDA is generated from a corpus $\mathbf{w}=\{w_{dsn}\}$ by

$$\prod_{d=1}^{M}\int p(\theta_d\,|\,\alpha)\prod_{s=1}^{S}\sum_{l=1}^{L}p(\eta_{ds}=l\,|\,\theta_d)\times\left[\prod_{n=1}^{N}\sum_{k=1}^{K}p(w_{dsn}\,|\,z_{dsn}=k,\beta)\,p(z_{dsn}=k\,|\,\mu,\eta_{ds})\right]d\theta_d\,. \tag{6}$$

Importantly, the additional theme variable is incorporated in SLDA model to characterize the latent thematic information at sentence level. Such an extension is crucial for document summarization.

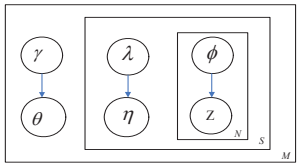

Figure 3: Graphical model for variational SLDA.

3.3 Inference in SLDA

Again, the direct optimization of (6) with respect to SLDA parameters $\{\alpha,\beta,\mu\}$ is intractable. We apply the variational inference and perform an approximate optimization of marginal likelihood through maximizing its lower bound [9]. Convergence property is assured because the bound is convex with respect to the factored variational distributions. We simplify the inference in SLDA by assuming the conditional independence among latent variables $\{\theta,\eta,\mathbf{z}\}$ as illustrated in Figure 3. The variational parameters $\{\gamma,\lambda,\phi\}$ act as the hyperparameters of variational distribution $q(\theta,\eta,\mathbf{z}\,|\,\gamma,\lambda,\phi)$. The inferred model $q(\theta,\eta,\mathbf{z}\,|\,\gamma,\lambda,\phi)$ approximates the posterior distribution $p(\theta,\eta,\mathbf{z}\,|\,\mathbf{w},\alpha,\beta,\mu)$ with the smallest KL divergence. The lower bound of (6) is expanded by factorizing *p* and *q* functions in a form of

$$\begin{aligned}L(\gamma,\lambda,\phi;\alpha,\beta,\mu)&=E_q[\ln p(\mathbf{w}\,|\,\mathbf{z},\beta)]+E_q[\ln p(\mathbf{z}\,|\,\eta,\mu)]\\&+E_q[\ln p(\eta\,|\,\theta)]+E_q[\ln p(\theta\,|\,\alpha)]-E_q[\ln q(\mathbf{z}\,|\,\phi)]\\&-E_q[\ln q(\eta\,|\,\lambda)]-E_q[\ln q(\theta\,|\,\gamma)].\end{aligned} \tag{7}$$

We derive the optimal variational parameters by

$$\phi_{slnk}\propto\mu_{lk}\beta_{kw_n} \tag{8}$$

$$\lambda_{sl}\propto\sum_{n=1}^{N}\sum_{k=1}^{K}\phi_{slnk}\,\exp\left[\Psi(\gamma_l)-\Psi(\sum_{j=1}^{L}\gamma_j)\right] \tag{9}$$

$$\gamma_l=(\alpha_l+\sum_{s=1}^{S}\lambda_{sl}) \tag{10}$$

and SLDA parameters by

$$\beta_{kw_n}\propto\sum_{d=1}^{M}\sum_{s=1}^{S}\sum_{l=1}^{L}\sum_{n=1}^{N}\lambda_{sl}\phi_{slnk}w_{dsn} \tag{11}$$

$$\mu_{lk}\propto\sum_{d=1}^{M}\sum_{s=1}^{S}\sum_{n=1}^{N}\phi_{dslnk}\lambda_{sl} \tag{12}$$

$$\alpha^{t+1}=\alpha^{t}-H_{\mathrm{SLDA}}(\alpha^{t})^{-1}g_{\mathrm{SLDA}}(\alpha^{t})\,. \tag{13}$$

where $H_{\mathrm{SLDA}}(\cdot)$ and $g_{\mathrm{SLDA}}(\cdot)$ denote the Hessian matrix and gradient vector of the lower bound of SLDA with respect to α, respectively. Owing to the unseen variables $\{\eta,\mathbf{z}\}$ in model inference, the variational Bayesian EM algorithm is applied. Such an iterative expectation and maximization steps shall converge to achieve the local optimum. The variational parameters γ,λ,ϕ are estimated in the first stage and the SLDA parameters α,μ,β are updated in the second stage. Different from LDA, SLDA is designed for exploring delicate structure in text documents, and works for the sentence selection in summarization procedure. The hierarchical information is incorporated into the estimated parameters in different levels. The physical meaning of variational parameters (8)-(10) is interpreted as follows. The parameter ϕ collects topic information for words, and is grouped in sentence level. Parameter λ is accumulated with the sentence and theme dependent parameter ϕ. Parameter γ absorbs the document-level information from the corresponding sentences by incorporating λ with theme labels. Notably, SLDA parameters α,μ,β are merged into variational parameters γ,λ,ϕ. The whole model is inferred and resulted in the sentence-based and document-based language models given by

$$p(w_n\,|\,s)=\sum_{l=1}^{L}\lambda_{sl}\sum_{k=1}^{K}\mu_{lk}\beta_{kw_n} \tag{14}$$

$$p(w_n\,|\,d)=\sum_{l=1}^{L}\gamma_l\sum_{k=1}^{K}\mu_{lk}\beta_{kw_n}\,. \tag{15}$$

These language models are finally used to calculate KL divergence for sentence selection. The proposed method can be applied to query-based summarization by finding the rank list of sentences according to the query likelihood calculated by SLDA parameters.

4. EXPERIMENTS

4.1 Experimental setup

In the experiments, we used DUC 2005 corpus (http://duc.nist.gov/) where each document contained news articles from 50 topics. There were 25-50 news articles in a topic. The number of total sentences in this corpus was 37787 and the vocabulary size was 22613. This database provided the reference summaries which were manually written for evaluation of multi-document summarization, and also provided the query sentences for evaluation of query-based summarization. The NIST evaluation tool, called ROUGE (Recall-Oriented Understudy for Gisting Evaluation) at http://haydn.isi.edu/ROUGE, was adopted. ROUGE-N was used to measure the matched *n*-gram between reference and

automatic summaries, and ROUGE-L was used to calculate the longest common subsequence between two text datasets. The automatic summary for DUC was limited to 250 words at most. In the experiments, VSM [6] and LDA were implemented. For comparison, we carried out the language model (LM) summarization where KL divergence between unigram models of sentence and document was evaluated. In LDA and SLDA models, we adopted the number of topics as K=20, 50 and 100 and the number of themes as L=50 and 100.

Table 1: Precisions of LDA and SLDA using ROUGE-1 with different numbers of latent topics and themes

	K=20		K=50		K=100	
LDA	0.314		0.326		0.318	
	L=50	L=100	L=50	L=100	L=50	L=100
SLDA	0.352	0.378	0.376	0.389	0.369	0.384

Table 2: Comparison of recall (R), precision (P) and F-measure (F) for VSM, LM, LDA and SLDA

	ROUGE-1			ROUGE-2			ROUGE-L		
	R	P	F	R	P	F	R	P	F
VSM	0.3238	0.2958	0.3089	0.0440	0.0401	0.0419	0.2982	0.2725	0.2845
LM	0.3437	0.2863	0.3114	0.0472	0.0394	0.0428	0.3017	0.2513	0.2734
LDA	0.3141	0.3261	0.3170	0.0463	0.0471	0.0466	0.2844	0.2909	0.2869
SLDA	0.3372	0.3897	0.3580	0.0739	0.060	0.0600	0.2982	0.3395	0.3164

4.2 Experimental results

First, we compare the precisions of LDA and SLDA in Table 1. The best result was obtained by SLDA with K=50 and L=100. In Table 2, we show the summarization results of VSM, LM, LDA with K=50, and SLDA with K=50 and L=100. Here, ROUGE-1 and ROUGE-2 mean the evaluation of selected sentences by unigram and bigram schemes, respectively. The measures of recall, precision and F-measure were reported. In this set of experiments, SLDA consistently outperformed VSM and LDA in terms of different evaluation measures and tools. The improvement with unigram evaluation was better than that with bigram evaluation. It is because that LDA or SLDA are unigram-based models. We could improve ROUGE-2 performance if the latent Dirichlet learning is applied to bigram-based LDA or SLDA. Also, the performance of LDA was not as good as VSM and LM because the data of building sentence model using LDA was too sparse. However, such phenomenon does not exist in SLDA since SLDA efficiently calculates the statistics from word, sentence and document levels, and so the amount of text data is sufficient to build the whole model in one learning epoch. The statistics in different levels are coupled and used to generate the language models of document and individual sentences. SLDA achieved the best performance among these methods. In addition, we can determine the numbers of distribution parameters of VSM, LDA and SLDA as SV, $2(K+KV)$ and $L+LK+KV$, respectively. In current experimental setup, SLDA and LDA are comparable in terms of parameter size, but the computation cost of SLDA is considerably high.

5. CONCLUSION

We have presented a new hierarchical model to characterize the structure of documents, sentences and words in a text corpus. The SLDA model was built to compensate the weakness of small sample size in sentence modeling using traditional LDA. The Bayesian variational inference method was applied to solve the parameter inference problem. The robustness of the sentence model estimated from a set of documents was assured in the proposed SLDA-based summarization. This method delicately represented the structure of text corpus and experimentally worked better than the vector space summarization in terms of different evaluation measures. In the future, we will explore alternative inference solution to build SLDA model and work for reducing the computation cost. The experiments on spoken document summarization will be conducted.

6. REFERENCES

[1] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet allocation", *Journal of Machine Learning Research*, vol. 3, no. 5, pp. 993-1022, 2003.

[2] R. Brandow, K. Mitze and L. F. Rau, "Automatic condensation of electronic publications by sentence selection", *Information Processing and Management*, vol. 31, no. 5, pp. 675-685, 1995.

[3] J.-T. Chien and M.-S. Wu, "Adaptive Bayesian latent semantic analysis", *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 198-207, 2008.

[4] J.-T. Chien and C.-H. Chueh, "Latent Dirichlet language model for speech recognition", in *Proc. of IEEE Workshop on Spoken Language Technology*, pp. 201-204, 2008.

[5] S. Furui, T. Kikuchi, Y. Shinnaka, C. Hori, "Speech-to-text and speech-to-speech summarization of spontaneous speech", *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 4, pp. 401-408, 2004.

[6] Y. Gong, and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis", *in Proc. of ACM SIGIR*, pp. 19-25, 2001.

[7] M. Hirohata, Y. Shinnaka, K. Iwano and S. Furui, "Sentence-extractive automatic speech summarization and evaluation techniques", *Speech Communication*, vol. 48, no. 9, pp. 1151-1161, 2006.

[8] T. Hofmann, "Probabilistic latent semantic indexing", in *Proc. of ACM SIGIR*, pp. 35-44, 1999.

[9] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola and L Saul, "An introduction to variational methods for graphical models", *Machine Learning*, vol. 37, pp. 183-233, 1999.

[10] D. R. Radev, H. Jing, M. Sty and D. Tam, "Centroid-based summarization of multiple documents", *Information Processing and Management*, vol. 40, no. 6, pp. 919-938, 2004.

[11] M. M. Shafiei and E. E. Milios, "Latent Dirichlet co-clustering", in *Proc. of International Conference on Data Mining*, pp. 542-551, 2006.