A Novel Feature-based Bayesian Model for Query Focused Multi-document Summarization

Jiwei Li
School of Computer Science
Carnegie Mellon University
bdlijiwei@gmail.com

Sujian Li
Laboratory of Computational Linguistics
Peking University
lisujian@pku.edu.cn

Abstract

Supervised learning methods and LDA based topic model have been successfully applied in the field of multi-document summarization. In this paper, we propose a novel supervised approach that can incorporate rich sentence features into Bayesian topic models in a principled way, thus taking advantages of both topic model and feature based supervised learning methods. Experimental results on DUC2007, TAC2008 and TAC2009 demonstrate the effectiveness of our approach.

1 Introduction

Query-focused multi-document summarization (Nenkova et al., 2006; Wan et al., 2007; Ouyang et al., 2010) can facilitate users to grasp the main idea of documents. In query-focused summarization, a specific topic description, such as a query, which expresses the most important topic information is proposed before the document collection, and a summary would be generated according to the given topic.

Supervised models have been widely used in summarization (Li, et al., 2009, Shen et al., 2007, Ouyang et al., 2010). Supervised models usually regard summarization as a classification or regression problem and use various sentence features to build a classifier based on labeled negative or positive samples. However, existing supervised approaches seldom exploit the intrinsic structure among sentences. This disadvantage usually gives rise to serious problems such as unbalance and low recall in summaries.

Recently, LDA-based (Blei et al., 2003) Bayesian topic models have widely been applied in multi-document summarization in that Bayesian approaches can offer clear and rigorous probabilistic interpretations for summaries(Daume and Marcu, 2006; Haghighi and Vanderwende, 2009; Jin et al., 2010; Mason and Charniak, 2011; Delort and Alfonseca, 2012). Exiting Bayesian approaches label sentences or words with topics and sentences which are closely related with query or can highly generalize documents are selected into summaries. However, LDA topic model suffers from the intrinsic disadvantages that it only uses word frequency for topic modeling and can not use useful text features such as position, word order etc (Zhu and Xing, 2010). For example, the first sentence in a document may be more important for summary since it is more likely to give a global generalization about the document. It is hard for LDA model to consider such information, making useful information lost.

It naturally comes to our minds that we can improve summarization performance by making full use of both useful text features and the latent semantic structures from by LDA topic model. One related work is from Celikyilmaz and Hakkani-Tur (2010). They built a hierarchical topic model called Hybhsum based on LDA for topic discovery and assumed this model can produce appropriate scores for sentence evaluation. Then the scores are used for tuning the weights of various features that helpful for summary generation. Their work made a good step of combining topic model with feature based supervised learning. However, what their approach confuses us is that whether a topic model only based on word frequency is good enough to generate an appropriate sentence score for regression. Actually, how to incorporate features into LDA topic model has been a open problem. Supervised topic models such as sLDA(Blei and MacAuliffe 2007) give us some inspiration. In sLDA, each document is associated with a labeled feature and sLDA can integrate such feature into LDA for topic modeling in a prin-

Transactions of the Association for Computational Linguistics, 1 (2013) 89–98. Action Editor: Noah Smith.
Submitted 12/2012; Published 5/2013. ©2013 Association for Computational Linguistics.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00212/1566641/tacl_a_00212.pdf by guest on 07 December 2024

cipled way.

With reference to the work of supervised LDA models, in this paper, we propose a novel sentence feature based Bayesian model S-sLDA for multi-document summarization. Our approach can naturally combine feature based supervised methods and topic models. The most important and challenging problem in our model is the tuning of feature weights. To solve this problem, we transform the problem of finding optimum feature weights into an optimization algorithm and learn these weights in a supervised way. A set of experiments are conducted based on the benchmark data of DUC2007, TAC2008 and TAC2009, and experimental results show the effectiveness of our model.

The rest of the paper is organized as follows. Section 2 describes some background and related works. Section 3 describes our details of S-sLDA model. Section 4 demonstrates details of our approaches, including learning, inference and summary generation. Section 5 provides experiments results and Section 6 concludes the paper.

2 Related Work

A variety of approaches have been proposed for query-focused multi-document summarizations such as unsupervised (semi-supervised) approaches, supervised approaches, and Bayesian approaches.

Unsupervised (semi-supervised) approaches such as Lexrank (Erkan and Radex, 2004), manifold (Wan et al., 2007) treat summarization as a graph-based ranking problem. The relatedness between the query and each sentence is achieved by imposing querys influence on each sentence along with the propagation of graph. Most supervised approaches regard summarization task as a sentence level two class classification problem. Supervised machine learning methods such as Support Vector Machine(SVM) (Li, et al., 2009), Maximum Entropy (Osborne, 2002) , Conditional Random Field (Shen et al., 2007) and regression models (Ouyang et al., 2010) have been adopted to leverage the rich sentence features for summarization.

Recently, Bayesian topic models have shown their power in summarization for its clear probabilistic interpretation. Daume and Marcu (2006) proposed Bayesum model for sentence extraction based on query expansion concept in information retrieval. Haghighi and Vanderwende (2009) proposed topicsum and hiersum which use a LDA-like topic model and assign each sentence a distribution over background topic, doc-specific topic and content topics. Celikyilmaz and Hakkani-Tur (2010) made a good step in combining topic model with supervised feature based regression for sentence scoring in summarization. In their model, the score of training sentences are firstly got through a novel hierarchical topic model. Then a featured based support vector regression (SVR) is used for sentence score prediction. The problem of Celikyilmaz and Hakkani-Turs model is that topic model and feature based regression are two separate processes and the score of training sentences may be biased because their topic model only consider word frequency and fail to consider other important features. Supervised feature based topic models have been proposed in recent years to incorporate different kinds of features into LDA model. Blei (2007) proposed sLDA for document response pairs and Daniel et al. (2009) proposed Labeled LDA by defining a one to one correspondence between latent topic and user tags. Zhu and Xing (2010) proposed conditional topic random field (CTRF) which addresses feature and independent limitation in LDA.

3 Model description

3.1 LDA and sLDA

The hierarchical Bayesian LDA (Blei et al., 2003) models the probability of a corpus on hidden topics as shown in Figure 1(a). Let K be the number of topics , M be the number of documents in the corpus and V be vocabulary size. The topic distribution of each document θ_m is drawn from a prior Dirichlet distribution $Dir(\alpha)$, and each document word w_{mn} is sampled from a topic-word distribution ϕ^z specified by a drawn from the topic-document distribution θ_m. β is a $K \times M$ dimensional matrix and each β_k is a distribution over the V terms. The generating procedure of LDA is illustrated in Figure 2. θ_m is a mixture proportion over topics of document m and z_{mn} is a K dimensional variable that presents the topic assignment distribution of different words.

Supervised LDA (sLDA) (Blei and McAuliffe 2007) is a document feature based model and intro-

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00212/1566641/tacl_a_00212.pdf by guest on 07 December 2024

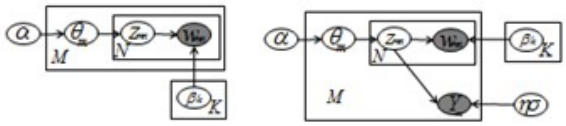

Figure 1: Graphical models for (a) LDA model and (b) sLDA model.

1. Draw a document proportion vector $\theta_m|\alpha \sim Dir(\alpha)$
2. For each word in m
   (a)draw topic assignment $z_{mn}|\theta \sim Multi(\theta_{z_{mn}})$
   (b)draw word $w_{mn}|z_{mn}, \beta \sim Multi(\beta_{z_{mn}})$

Figure 2: Generation process for LDA

duces a response variable to each document for topic discovering, as shown in Figure 1(b). In the generative procedure of sLDA, the document pairwise label is draw from $y|\overrightarrow{z_m}, \eta, \delta^2 \sim p(y|\overrightarrow{z_m}, \eta, \delta^2)$, where $\overrightarrow{z_m} = \frac{1}{N}\sum_{n=1}^{N} z_{m,n}$.

3.2 Problem Formulation

Here we firstly give a standard formulation of the task. Let K be the number of topics, V be the vocabulary size and M be the number of documents. Each document D_m is represented with a collection of sentence $D_m = \{S_s\}_{s=1}^{s=N_m}$ where N_m denotes the number of sentences in m^{th} document. Each sentence is represented with a collection of words $\{w_{msn}\}_{n=1}^{n=N_{ms}}$ where N_{ms} denotes the number of words in current sentence. $\overrightarrow{Y_{ms}}$ denotes the feature vector of current sentence and we assume that these features are independent.

3.3 S-sLDA

z_{ms} is the hidden variable indicating the topic of current sentence. In S-sLDA, we make an assumption that words in the same sentence are generated from the same topic which was proposed by Gruber (2007). z_{msn} denotes the topic assignment of current word. According to our assumption, $z_{msn} = z_{ms}$ for any $n \in [1, N_{ms}]$. The generative approach of S-sLDA is shown in Figure 3 and Figure 4. We can see that the generative process involves not only the words within current sentence, but also a series of sentence features. The mixture weights over features in S-sLDA are defined with a generalized linear model (GLM).

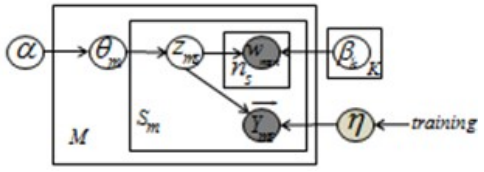

Figure 3: Graph model for S-sLDA model

1. Draw a document proportion vector $\theta_m|\alpha \sim Dir(\alpha)$
2. For each sentence in m
   (a)draw topic assignment $z_{ms}|\theta \sim Multi(\theta_{z_{mn}})$
   (b)draw feature vector $\overrightarrow{Y_{ms}}|z_{ms}, \eta \sim p(\overrightarrow{Y_{ms}}|z_{ms}, \eta)$
   (c)for each word w_{msn} in current sentence
   draw $w_{msn}|z_{ms}, \beta \sim Multi(\beta_{z_{ms}})$

Figure 4: generation process for S-sLDA

$$p(\overrightarrow{Y_{ms}}|z_{ms}, \eta) = \frac{exp(z_{ms}^T \eta)\overrightarrow{Y_{ms}}}{\sum_{z_{ms}} exp(z_{ms}^T \eta)\overrightarrow{Y_{ms}}} \quad (1)$$

Here we assume that each sentence has T features and $\overrightarrow{Y_{ms}}$ is a $T \times 1$ dimensional vector. η is a $K \times T$ weight matrix of each feature upon topics, which largely controls the feature generation procedure. Unlike s-LDA where η is a latent variable estimated from the maximum likelihood estimation algorithm, in S-sLDA the value of η is trained through a supervised algorithm which will be illustrated in detail in Section 3.

3.4 Posterior Inference and Estimation

Given a document and labels for each sentence, the posterior distribution of the latent variables is:

$$p(\theta, z_{1:N}|w_{1:N}, Y, \alpha, \beta_{1:K}, \eta) = \frac{\prod_m p(\theta_m|\alpha)\prod_s [p(z_{ms}|\theta_m)p(\overrightarrow{Y_{ms}}|z_{ms}, \eta)\prod_n p(w_{msn}|z_{msn}, \beta_{z_{msn}})}{\int d\theta p(\theta_m|\alpha)\sum_z \prod_s [p(z_{ms}|\theta_m)p(\overrightarrow{Y_{ms}}|z_{ms}, \eta)\prod_n p(w_{msn}|\beta_{z_{msn}})} \quad (2)$$

Eqn. (2) cannot be efficiently computed. By applying the Jensens inequality, we obtain a lower bound of the log likelihood of document $p(\theta, z_{1:N}|w_{1:N}, \overrightarrow{Y_{ms}}, \alpha, \beta_{1:K}, \eta) \geq L$, where

$$L = \sum_{ms} E[logP(z_m s|\theta)] + \sum_{ms} E[logP(\overrightarrow{Y_{ms}}|z_{ms}, \eta)] + \sum_m E[logP(\theta|\alpha)] + \sum_{msn} E[logP(w_{msn}|z_{ms}, \beta)] + H(q) \quad (3)$$

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00212/1566641/tacl_a_00212.pdf by guest on 07 December 2024

where $H(q) = -E[logq]$ and it is the entropy of variational distribution q is defined as

$$q(\theta, z|\gamma, \phi) = \prod_{mk} q(\theta_m|\gamma) \prod_{sn} q(z_{msn}|\phi_{ms}) \quad (4)$$

here γ a K-dimensional Dirichlet parameter vector and multinomial parameters. The first, third and forth terms of Eqn. (3) are identical to the corresponding terms for unsupervised LDA (Blei et al., 2003). The second term is the expectation of log probability of features given the latent topic assignments.

$$E[logP(\overrightarrow{Y_{ms}}|z_{ms}, \eta)] = E(z_{ms})^T \eta \overrightarrow{Y_{ms}} - log \sum_{z_{ms}} exp(z_{ms}^T \eta \overrightarrow{Y_{ms}}) \quad (5)$$

where $E(z_{ms})^T$ is a $1 \times K$ dimensional vector $[\phi_{msk}]_{k=1}^{k=K}$. The Bayes estimation for S-sLDA model can be got via a variational EM algorithm. In EM procedure, the lower bound is firstly minimized with respect to γ and ϕ, and then minimized with α and β by fixing γ and ϕ.

E-step:

The updating of Dirichlet parameter γ is identical to that of unsupervised LDA, and does not involve feature vector $\overrightarrow{Y_{ms}}$.

$$\gamma_m^{new} \leftarrow \alpha + \sum_{s \in m} \phi_s \quad (6)$$

$$\phi_{sk}^{new} \propto exp\{E[log\theta_m|\gamma] + \sum_{n=1}^{N_{ms}} E[log(w_{msn}|\beta_{1:K})] + \sum_{t=1}^{T} \eta_{kt} Y_{st}\} = exp[\Psi(\gamma_{mk}) - \Psi(\sum_{k=1}^{K} \gamma_{mk}) + \sum_{t=1}^{T} \eta_{kt} Y_{st}] \quad (7)$$

where $\Psi(\cdot)$ denotes the log Γ function. m_s denotes the document that current sentence comes from and Y_{st} denotes the t^{th} feature of sentence s.

M-step:

The M-step for updating β is the same as the procedure in unsupervised LDA, where the probability of a word generated from a topic is proportional to the number of times this word assigned to the topic.

$$\beta_{kw}^{new} = \sum_{m=1}^{M} \sum_{s=1}^{N_m} \sum_{n=1}^{N_{ms}} 1(w_{msn} = w) \phi_{ms}^k \quad (8)$$

4 Our Approach

4.1 Learning

In this subsection, we describe how we learn the feature weight η in a supervised way. The learning process of η is a supervised algorithm combined with variational inference of S-sLDA. Given a topic description Q[1] and a collection of training sentences S from related documents, human assessors assign a score $v(v = -2, -1, 0, 1, 1)$ to each sentence in S. The score is an integer between -2 (the least desired summary sentences) and $+2$ (the most desired summary sentences), and score 0 denotes neutral attitude. $O_v = \{o_{v1}, o_{v2}, ..., v_{vk}\}(v = -2, -1, 0, 1, 2)$ is the set containing sentences with score v. Let ϕ_{Qk} denote the probability that query is generated from topic k. Since query does not belong to any document, we use the following strategy to leverage ϕ_{Qk}

$$\phi_{Qk} = \prod_{w \in Q} \beta_{kw} \cdot \frac{1}{M} \sum_{m=1}^{M} exp[\Psi(\gamma_{mk}) - \Psi(\sum_{k=1}^{K} \gamma_{mk})] \quad (9)$$

In Equ.(9), $\prod_{w \in Q} \beta_{kw}$ denotes the probability that all terms in query are generated from topic k and $\frac{1}{M} \sum_{m=1}^{M} exp[\Psi(\gamma_{mk}) - \Psi(\sum_{k=1}^{K} \gamma_{mk})]$ can be seen as the average probability that all documents in the corpus are talking about topic k. Eqn. (9) is based on the assumption that query topic is relevant to the main topic discussed by the document corpus. This is a reasonable assumption and most previous LDA summarization models are based on similar assumptions.

Next, we define $\phi_{O_v,k}$ for sentence set O_v, which can be interpreted as the probability that all sentences in collection O_v are generated from topic k.

$$\phi_{O_v,k} = \frac{1}{|O_v|} \sum_{s \in O_v} \phi_{sk}, k \in [1, K], v \in [-2, 2] \quad (10)$$

$|O_v|$ denotes the number of sentences in set O_v. Inspired by the idea that desired summary sentences would be more semantically related with the query, we transform problem of finding optimum η to the following optimization problem:

$$min_\eta L(\eta) = \sum_{v=-2}^{v=2} v \cdot KL(O_v||Q); \quad \sum_{t=1}^{T} \eta_{kt} = 1 \quad (11)$$

[1]We select multiple queries and their related sentences for training

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00212/1566641/tacl_a_00212.pdf by guest on 07 December 2024

where $KL(O_v||Q)$ is the Kullback-Leibler divergence between the topic and sentence set O_v as shown in Eqn.(12).

$$KL(O_v||Q) = \sum_{k=1}^{K} \phi_{O_v k} log \frac{\phi_{O_v k}}{\phi_{Qk}} \tag{12}$$

In Eqn. (11), we can see that O_2, which contain desirable sentences, would be given the largest penalty for its KL divergence from Query. The case is just opposite for undesired set.

Our idea is to incorporate the minimization process of Eqn.(11) into variational inference process of S-sLDA model. Here we perform gradient based optimization method to minimize Eqn.(11). Firstly, we derive the gradient of $L(\eta)$ with respect to η.

$$\frac{\partial L(\eta)}{\eta_{xy}} = \sum_{v=-2}^{v=2} v \cdot \frac{\partial KL(Q_v||Q)}{\partial \eta_{xy}} \tag{13}$$

$$\frac{\partial KL(Q_v||Q)}{\partial \eta_{xy}} = \sum_{k=1}^{K} \frac{1}{|Q_v|}(1 + log\frac{\sum_{s\in Q_v}}{|Q_v|}) \sum_{s\in Q_v} \frac{\partial \phi_{sk}}{\partial \eta_{xy}}$$
$$- \sum_{k=1}^{K} \frac{1}{|Q_v|} \sum_{s\in Q_v} \frac{\partial Q_{sk}}{\eta_{xy}} - \sum_{k=1}^{K} \frac{1}{Q_v} \frac{\sum_{s\in Q_v \phi_{sk}}}{\phi_{Qk}} \frac{\partial \phi_{sk}}{\partial \eta_{xy}} \tag{14}$$

For simplification, we regard β and γ as constant during updating process of η, so $\frac{\partial \phi_{Qk}}{\partial \eta_{xy}} = 0$.[2] We can further get first derivative for each labeled sentence.

$$\frac{\partial \phi_{sk}}{\eta_{xy}} \propto \begin{cases} Y_{sy} exp[\Psi(\gamma_{m_s i}) - \Psi(\sum_{k=1}^{K} \gamma_{m_s k}) + \sum_{t=1}^{T} \eta_{kt} Y_{sy}] & \\ \times \prod_{w\in s} \beta_{kw} & if\ k = x \\ 0 & if\ k \neq x \end{cases} \tag{15}$$

4.2 Feature Space

Lots of features have been proven to be useful for summarization (Louis et al., 2010). Here we discuss several types of features which are adopted in S-sLDA model. The feature values are either binary or normalized to the interval [0,1]. The following features are used in S-sLDA:

Cosine Similarity with query: Cosine similarity is based on the tf-idf value of terms.

[2]This is reasonable because the influence of γ and β have been embodied in ϕ during each iteration.

Local Inner-document Degree Order: Local Inner document Degree Order is a binary feature which indicates whether Inner-document Degree (IDD) of sentence s is the largest among its neighbors. IDD means the edge number between s and other sentences in the same document.

Document Specific Word: 1 if a sentence contains document specific word, 0 otherwise.

Average Unigram Probability (Nenkova and Vanderwende, 2005; Celikyilmaz and Hakkani-Tur 2010): As for sentence s, $p(s) = \sum_{w\in s} \frac{1}{|s|} p_D(w)$, where $p_D(w)$ is the observed unigram probability in document collection.

In addition, we also use the commonly used features including **sentence position**, **paragraph position**, **sentence length** and **sentence bigram frequency**.

E-step
  initialize $\phi^0_{sk} := 1/K$ for all i and s.
  initialize $\gamma_{mi} := \alpha_{mi} + N)m/K$ for all i.
  initialize $\eta_{kt} = 0$ for all k and t.
  while not convergence
  for $m = 1 : M$
    update γ^{t+1}_m according to Eqn.(6)
    for $s = 1 : N_m$
      for $k = 1 : K$
        update ϕ^{t+1}_{sk} according to Eqn.(7)
      normalize the sum of ϕ^{t+1}_{sk} to 1.
Minimize $L(\eta)$ according to Eqn.(11)-(15).
M-step:
  update β according to Eqn.(8)

Figure 5: Learning process of η in S-sLDA

4.3 Sentence Selection Strategy

Next we explain our sentence selection strategy. According to our intuition that the desired summary should have a small KL divergence with query, we propose a function to score a set of sentences Sum. We use a decreasing logistic function $\zeta(x) = 1/(1+e^x)$ to refine the score to the range of (0,1).

$$Score(Sum) = \zeta(KL(sum||Q)) \tag{16}$$

Let $Sum^\star$ denote the optimum update summary. We can get $Sum^\star$ by maximizing the scoring function.

$$Sum^\star = \underset{Sum\in S \&\& words(Sum)\leq L}{\arg\max} Score(Sum) \tag{17}$$

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00212/1566641/tacl_a_00212.pdf by guest on 07 December 2024

1. Learning: Given labeled set O_v, learn the feature weight vector η using algorithm in Figure 5.
2. Given new data set and η, use algorithm in section 3.3 for inference. (The only difference between this step and step (1) is that in this step we do not need minimize $L(\eta)$.
3. Select sentences for summarization from algorithm in Figure 6.

Figure 6: Summarization Generation by S-sLDA.

A greedy algorithm is applied by adding sentence one by one to obtain $Sum^{\star}$. We use G to denote the sentence set containing selected sentences. The algorithm first initializes G to Φ and X to SU. During each iteration, we select one sentence from X which maximize $Score(s_m \cup G)$. To avoid topic redundancy in the summary, we also revise the MMR strategy (Goldstein et al., 1999; Ouyang et al., 2007) in the process of sentence selection. For each s_m, we compute the semantic similarity between s_m and each sentence s_t in set Y in Eqn.(18).

$$cos-sem(s_m, s_t) = \frac{\sum_k \phi_{s_m k}\phi_{s_t k}}{\sqrt{\sum_k \phi^2_{s_m k}}\sqrt{\sum_k \phi^2_{s_t k}}} \quad (18)$$

We need to assure that the value of semantic similarity between two sentences is less than Th_{sem}. The whole procedure for summarization using S-sLDA model is illustrated in Figure 6. Th_{sem} is set to 0.5 in the experiments.

5 Experiments

5.1 Experiments Set-up

The query-focused multi-document summarization task defined in DUC[3](Document Understanding Conference) and TAC[4](Text Analysis Conference) evaluations requires generating a concise and well organized summary for a collection of related news documents according to a given query which describes the users information need. The query usually consists of a title and one or more narrative/question sentences. The system-generated summaries for DUC and TAC are respectively limited to 250 words and 100 words. Our experiment data is composed of DUC 2007, TAC[5] 2008 and TAC 2009 data which have 45, 48 and 44 collections respectively. In our experiments, DUC 2007 data is used as training data and TAC (2008-2009) data is used as the test data.

Stop-words in both documents and queries are removed using a stop-word list of 598 words, and the remaining words are stemmed by Porter Stemmer[6]. As for the automatic evaluation of summarization, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measures, including ROUGE-1, ROUGE-2, and ROUGE-SU4[7] and their corresponding 95% confidence intervals, are used to evaluate the performance of the summaries. In order to obtain a more comprehensive measure of summary quality, we also conduct manual evaluation on TAC data with reference to (Haghighi and Vanderwende, 2009; Celikyilmaz and Hakkani-Tur, 2011; Delort and Alfonseca, 2011).

5.2 Comparison with other Bayesian models

In this subsection, we compare our model with the following Bayesian baselines:

KL-sum: It is developed by Haghighi and Vanderwende (Lin et al., 2006) by using a KL-divergence based sentence selection strategy.

$$KL(P_s||Q_d) = \sum_w P(w) log \frac{P(w)}{Q(w)} \quad (19)$$

where P_s is the unigram distribution of candidate summary and Q_d denotes the unigram distribution of document collection. Sentences with higher ranking score is selected into the summary.

HierSum: A LDA based approach proposed by Haghighi and Vanderwende (2009), where unigram distribution is calculated from LDA topic model in Equ.(14).

Hybhsum: A supervised approach developed by Celikyilmaz and Hakkani-Tur (2010).

For fair comparison, baselines use the same preprecessing methods with our model and all sum-

[3]http://duc.nist.gov/.

[4]http://www.nist.gov/tac/.

[5]Here, we only use the docset-A data in TAC, since TAC data is composed of docset-A and docset-B data, and the docset-B data is mainly for the update summarization task.

[6]http://tartarus.org/ martin/PorterStemmer/.

[7]Jackknife scoring for ROUGE is used in order to compare with the human summaries.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00212/1566641/tacl_a_00212.pdf by guest on 07 December 2024

maries are truncated to the same length of 100 words. From Table 1 and Table 2, we can

Methods	ROUGE-1	ROUGE-2	ROUGE-SU4
Our approach	0.3724 (0.3660-0.3788)	0.1030 (0.0999-0.1061)	0.1342 (0.1290-0.1394)
Hybhsum	0.3703 (0.3600-0.3806)	0.1007 (0.0952-0.1059)	0.1314 (0.1241-0.1387)
HierSum	0.3613 (0.3374-0.3752)	0.0948 (0.0899-0.0998)	0.1278 (0.1197-0.1359)
KLsum	0.3504 (0.3411-0.3597)	0.0917 (0.0842-0.0992)	0.1234 (0.1155-0.1315)
StandLDA	0.3368 (0.3252-0.3386)	0.0797 (0.0758-0.0836)	0.1156 (0.1072-0.1240)

Table 1: Comparison of Bayesian models on TAC2008

Methods	ROUGE-1	ROUGE-2	ROUGE-SU4
Our approach	0.3903 (0.3819-0.3987)	0.1223 (0.1167-0.1279)	0.1488 (0.1446-0.1530)
Hybhsum	0.3824 (0.3686-0.3952)	0.1173 (0.1132-0.1214)	0.1436 (0.1358-0.1514)
HierSum	0.3706 (0.3624-0.3788)	0.1088 (0.0950-0.1144)	0.1386 (0.1312-0.1464)
KLsum	0.3619 (0.3510-0.3728)	0.0972 (0.0917-0.1047)	0.1299 (0.1213-0.1385)
StandLDA	0.3552 (0.3447-0.3657)	0.0847 (0.0813-0.0881)	0.1214 (0.1141-0.1286)

Table 2: Comparison of Bayesian models on TAC2009

see that among all the Bayesian baselines, Hybhsum achieves the best result. This further illustrates the advantages of combining topic model with supervised method. In Table 1, we can see that our S-sLDA model performs better than Hybhsum and the improvements are 3.4% and 3.7% with respect to ROUGE-2 and ROUGE-SU4 on TAC2008 data. The comparison can be extended to TAC2009 data as shown in Table 2: the performance of S-sLDA is above Hybhsum by 4.3% in ROUGE-2 and 5.1% in ROUGE-SU4. It is worth explaining that these achievements are significant, because in the TAC2008 evaluation, the performance of the top ranking systems are very close, i.e. the best system is only 4.2% above the 4th best system on ROUGE-2 and 1.2% on ROUGE-SU4.

5.3 Comparison with other baselines.

In this subsection, we compare our model with some widely used models in summarization.

Manifold: It is the one-layer graph based semi-supervised summarization approach developed by Wan et al.(2008). The graph is constructed only considering sentence relations using tf-idf and neglects topic information.

LexRank: Graph based summarization approach (Erkan and Radev, 2004), which is a revised version of famous web ranking algorithm PageRank. It is an unsupervised ranking algorithms compared with Manifold.

SVM: A supervised method - Support Vector Machine (SVM) (Vapnik 1995) which uses the same features as our approach.

MEAD: A centroid based summary algorithm by Radev et al. (2004). Cluster centroids in MEAD consists of words which are central not only to one article in a cluster, but to all the articles. Similarity is measure using tf-idf.

At the same time, we also present the top three participating systems with regard to ROUGE-2 on TAC2008 and TAC2009 for comparison, denoted as (denoted as SysRank 1st, 2nd and 3rd)(Gillick et al., 2008; Zhang et al., 2008; Gillick et al., 2009; Varma et al., 2009). The ROUGE scores of the top TAC system are directly provided by the TAC evaluation.

From Table 3 and Table 4, we can see that our approach outperforms the baselines in terms of ROUGE metrics consistently. When compared with the standard supervised method SVM, the relative improvements over the ROUGE-1, ROUGE-2 and ROUGE-SU4 scores are 4.3%, 13.1%, 8.3% respectively on TAC2008 and 7.2%, 14.9%, 14.3% on TAC2009. Our model is not as good as top participating systems on TAC2008 and TAC2009. But considering the fact that our model neither uses sentence compression algorithm nor leverage domain knowledge bases like Wikipedia or training data, such small difference in ROUGE scores is reasonable.

5.4 Manual Evaluations

In order to obtain a more accurate measure of summary quality for our S-sLDA model and Hybhsum, we performed a simple user study concerning the following aspects: (1) Overall quality: Which summary is better overall? (2) Focus: Which summary contains less irrelevant content? (3)Responsiveness: Which summary is more responsive to the query. (4) Non-Redundancy: Which summary is less redundant? 8 judges who specialize in NLP participated in the blind evaluation task. Evaluators are presented with two summaries generated by S-sLDA

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00212/1566641/tacl_a_00212.pdf by guest on 07 December 2024

Methods	ROUGE-1	ROUGE-2	ROUGE-SU4
Our approach	0.3724 (0.3660-0.3788)	0.1030 (0.0999-0.1061)	0.1342 (0.1290-0.1394)
SysRank 1^{st}	0.3742 (0.3639-0.3845)	0.1039 (0.0974-0.1104)	0.1364 (0.1285-0.1443)
SysRank 2^{nd}	0.3717 (0.3610-0.3824	0.0990 (0.0944-0.1038)	0.1326 (0.1269-0.1385)
SysRank 3^{rd}	0.3710 (0.3550-0.3849)	0.0977 (0.0920-0.1034)	0.1329 (0.1267-0.1391)
PageRank	0.3597 (0.3499-0.3695)	0.0879 (0.0809-0.0950)	0.1221 (0.1173-0.1269)
Manifold	0.3621 (0.3506-0.3736)	0.0931 (0.0868-0.0994)	0.1243 (0.1206-0.1280)
SVM	0.3588 (0.3489-0.3687)	0.0921 (0.0882-0.0960)	0.1258 (0.1204-0.1302)
MEAD	0.3558 (0.3489-0.3627)	0.0917 (0.0882-0.0952)	0.1226 (0.1174-0.1278)

Table 3: Comparison with baselines on TAC2008

Methods	ROUGE-1	ROUGE-2	ROUGE-SU4
Our approach	0.3903 (0.3819-0.3987)	0.1223 (0.1167-0.1279)	0.1488 (0.1446-0.1530)
SysRank 1st	0.3917 (0.3778-0.4057)	0.1218 (0.1122-0.1314)	0.1505 (0.1414-0.1596)
SysRank 2nd	0.3914 (0.3808-0.4020)	0.1212 (0.1147-0.1277)	0.1513 (0.1455-0.1571)
SysRank 3rd	0.3851 (0.3762-0.3932)	0.1084 (0.1025-0.1144)	0.1447 (0.1398-0.1496)
PageRank	0.3616 (0.3532-0.3700)	0.0849 (0.0802-0.0896)	0.1249 (0.1221-0.1277)
Manifold	0.3713 (0.3586-0.3841)	0.1014 (0.0950-0.1178)	0.1342 (0.1299-0.1385)
SVM	0.3649 (0.3536-0.3762)	0.1028 (0.0957-0.1099)	0.1319 (0.1258-0.1380)
MEAD	0.3601 (0.3536-0.3666)	0.1001 (0.0953-0.1049)	0.1287 (0.1228-0.1346)

Table 4: Comparison with baselines on TAC2009

and Hybhsum, as well as the four questions above. Then they need to answer which summary is better (tie). We randomly select 20 document collections from TAC 2008 data and randomly assign two summaries for each collection to three different evaluators to judge which model is better in each aspect.

As we can see from Table 5, the two models almost tie with respect to Non-redundancy, mainly because both models have used appropriate MMR strategies. But as for Overall quality, Focus and Responsiveness, S-sLDA model outputs Hybhsum based on t-test on 95% confidence level. Table 6 shows the example summaries generated respectively by two models for document collection D0803A-A in TAC2008, whose query is "Describe the coal mine accidents in China and actions taken". From table 6, we can see that each sentence in these two summaries is somewhat related to topics of coal mines in China. We also observe that the summary in Table 6(a) is better than that in Table 6(b), tending to select shorter sentences and provide more information. This is because, in S-sLDA model, topic modeling is determined simultaneously by various features including terms and other ones such as sentence length, sentence position and so on, which can contribute to summary quality. As we can see, in Table 6(b), sentences (3) and (5) provide some unimportant information such as "somebody said", though they contain some words which are related to topics about coal mines.

	Our(win)	Hybhsum(win)	Tie
Overall	37	14	9
Focus	32	18	10
Responsiveness	33	13	14
Non-redundancy	13	11	36

Table 5: Comparison with baselines on TAC2009

(1)China to close at least 4,000 coal mines this year: official (2)By Oct. 10 this year there had been 43 coal mine accidents that killed 10 or more people, (3)Officials had stakes in coal mines. (4)All the coal mines will be closed down this year. (5) In the first eight months, the death toll of coal mine accidents rose 8.5 percent last year. (6) The government has issued a series of regulations and measures to improve the coun.try's coal mine safety situation. (7)The mining safety technology and equipments have been sold to countries. (8)More than 6,000 miners died in accidents in China

(1) In the first eight months, the death toll of coal mine accidents across China rose 8.5 percent from the same period last year. (2)China will close down a number of ill-operated coal mines at the end of this month, said a work safety official here Monday. (3) Li Yizhong, director of the National Bureau of Production Safety Supervision and Administration, has said the collusion between mine owners and officials is to be condemned. (4)from January to September this year, 4,228 people were killed in 2,337 coal mine accidents. (5) Chen said officials who refused to register their stakes in coal mines within the required time

Table 6: Example summary text generated by systems (a)S-sLDA and (b) Hybhsum. (D0803A-A, TAC2008)

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00212/1566641/tacl_a_00212.pdf by guest on 07 December 2024

6 Conclusion

In this paper, we propose a novel supervised approach based on revised supervised topic model for query-focused multi document summarization. Our approach naturally combines Bayesian topic model with supervised method and enjoy the advantages of both models. Experiments on benchmark demonstrate good performance of our model.

Acknowledgments

This research work has been supported by NSFC grants (No.90920011 and No.61273278), National Key Technology R&D Program (No:2011BAH1B0403), and National High Technology $R\&D$ Program (No.2012AA011101). We also thank the three anonymous reviewers for their helpful comments. Corresponding author: Sujian Li.

References

David Blei and Jon McAuliffe. Supervised topic models. 2007. In *Neural Information Processing Systems*

David Blei, Andrew Ng and Micheal Jordan. Latent dirichlet allocation. In *The Journal of Machine Learning Research*, page: 993-1022.

Charles Broyden. 1965. A class of methods for solving nonlinear simultaneous equations. In *Math. Comp.* volume 19, page 577-593.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*.

Asli Celikyilmaz and Dilek Hakkani-Tur. 2010. A Hybrid hierarchical model for multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. page: 815-825

Jade Goldstein, Mark Kantrowitz, Vibhu Mittal and Jaime Carbonell. 1999. Summarizing Text Documents: Sentence Selection and Evaluation Metrics. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, page: 121-128.

Amit Grubber, Micheal Rosen-zvi and Yair Weiss. 2007. Hidden Topic Markov Model. *In Artificial Intelligence and Statistics.*

Hal Daume and Daniel Marcu H. 2006. Bayesian Query-Focused Summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, page 305-312.

Gune Erkan and Dragomir Radev. 2004. Lexrank: graph-based lexical centrality as salience in text summarization. In *J. Artif. Intell. Res. (JAIR)*, page 457-479.

Dan Gillick, Benoit Favre, Dilek Hakkani-Tur, *The ICSI Summarization System at TAC*, TAC 2008.

Dan Gillick, Benoit Favre, and Dilek Hakkani-Tur, Berndt Bohnet, Yang Liu, Shasha Xie. *The ICSI/UTD Summarization System at TAC 2009*. TAC 2009

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362370.

Feng Jin, Minlie Huang, and Xiaoyan Zhu. 2010. The summarization systems at tac 2010. In *Proceedings of the third Text Analysis Conference*, TAC-2010.

Liangda Li, Ke Zhou, Gui-Rong Xue, Hongyuan Zha and Yong Yu. 2009. Enhancing diversity, coverage and balance for summarization through structure learning. In *Proceedings of the 18th international conference on World wide web*, page 71-80.

Chin-Yew Lin, Guihong Gao, Jianfeng Gao and Jian-Yun Nie. 2006. An information-theoretic approach to automatic evaluation of summaries. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, page:462-470.

Annie Louis, Aravind Joshi, Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page:147-156.

Tengfei Ma, Xiaojun Wan. 2010. Multi-document summarization using minimum distortion, in *Proceedings of International Conference of Data Mining*. page 354363.

Rebecca Mason and Eugene Charniak. 2011. Extractive multi-document summaries should explicitly not contain document-specific content. In *proceedings of ACL HLT*, page:49-54.

Ani Nenkova and Lucy Vanderwende. The impact of frequency on summarization. In *Tech. Report MSR-TR-2005-101*, Microsoft Research, Redwood, Washington, 2005.

Ani Nenkova, Lucy Vanderwende and Kathleen McKeown. 2006. A compositional context sensitive multi-document summarizer: exploring the factors that inuence summarization. In *Proceedings of the 29th annual International ACM SIGIR Conference on Re-*

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00212/1566641/tacl_a_00212.pdf by guest on 07 December 2024

search and Development in Information Retrieval, page 573-580.

Miles Osborne. 2002. Using maximum entropy for sentence extraction. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, Volume 4 page:1-8.

Jahna Otterbacher, Gunes Erkan and Dragomir Radev. 2005. Using random walks for question-focused sentence retrieval. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, page 915-922

You Ouyang, Wenjie Li, Sujian Li and Qin Lua. 2011. Applying regression models to query-focused multi-document summarization. In *Information Processing and Management*, page 227-237.

You Ouyang, Sujian. Li, and Wenjie. Li. 2007, Developing learning strategies for topic-based summarization. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, page: 7986.

Daniel Ramage, David Hall, Ramesh Nallapati and Christopher Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Vol 1, page 248-256.

Dou She, Jian-Tao Sun, Hua Li, Qiang Yang and Zheng Chen. 2007. Document summarization using conditional random elds. In *Proceedings of International Joint Conference on Artificial Intelligence*, page: 28622867.

V. Varma, V. Bharat, S. Kovelamudi, P. Bysani, S. GSK, K. Kumar N, K. Reddy, N. Maganti , IIIT Hyderabad at *TAC 2009*. TAC2009

Xiaojun Wan and Jianwu Yang. 2008. Multi-document Summarization using cluster-based link analysis. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, page: 299-306.

Xiaojun Wan, Jianwu Yang and Jianguo Xiao. 2007. Manifold-ranking based topic-focused multi-document summarization. In *Proceedings of International Joint Conference on Artificial Intelligence*, page 2903-2908.

Furu Wei, Wenjie Li, Qin Lu and Yanxiang He. 2008. Exploiting Query-Sensitive Similarity for Graph-Based Query-Oriented Summarization. In *Proceedings of the 31st annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 283-290.

Jin Zhang, Xueqi Cheng, Hongbo Xu, Xiaolei Wang, Yiling Zeng. ICTCAS's ICTGrasper at TAC 2008: Summarizing Dynamic Information with Signature Terms Based Content Filtering, TAC 2008.

Dengzhong Zhou, Jason Weston, Arthur Gretton, Olivier Bousquet and Bernhard Schlkopf. 2003. Ranking on Data Manifolds. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*, page 169-176.

Jun Zhu and Eric Xing. 2010. Conditional Topic Random Fields. In *Proceedings of the 27th International Conference on Machine Learning.*

Xiaojin Zhu, Zoubin Ghahramani and John Lafferty. 2003. Semi-supervised Learning using Gaussian Fields and Harmonic Functions. In *Proceedings of International Conference of Machine Learning*, page: 912-919.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00212/1566641/tacl_a_00212.pdf by guest on 07 December 2024