Automatically Determining a Proper Length for Multi-document Summarization: A Bayesian Nonparametric Approach

Tengfei Ma and Hiroshi Nakagawa
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo
{matf@r., nakagawa@}dl.itc.u-tokyo.ac.jp

Abstract

Document summarization is an important task in the area of natural language processing, which aims to extract the most important information from a single document or a cluster of documents. In various summarization tasks, the summary length is manually defined. However, how to find the proper summary length is quite a problem; and keeping all summaries restricted to the same length is not always a good choice. It is obviously improper to generate summaries with the same length for two clusters of documents which contain quite different quantity of information. In this paper, we propose a Bayesian nonparametric model for multi-document summarization in order to automatically determine the proper lengths of summaries. Assuming that an original document can be reconstructed from its summary, we describe the ”reconstruction” by a Bayesian framework which selects sentences to form a good summary. Experimental results on DUC2004 data sets and some expanded data demonstrate the good quality of our summaries and the rationality of the length determination.

1 Introduction

Text summarization is the process of generating a short version of a given text to indicate its main topics. As the number of documents on the web exponentially increases, text summarization has attracted increasing attention, because it can help people get the most important information within a short time.

In most of the existing summarization systems, people need to first define a constant length to restrict all the output summaries. However, in many cases it is improper to require all summaries are of the same length. Take the multi-document summarization as an example, generating the summaries of the same length for a 5-document cluster and a 50-document cluster is intuitively improper. More specifically, consider two different clusters of documents: one cluster contains very similar articles which all focus on the same event at the same time; the other contains different steps of the event but each step has its own topics. The former cluster may need only one or two sentences to explain its information, while the latter needs to include more.

Research on summary length dates back in the late 90s. Goldstein et al. (1999) studied the characteristics of a good summary (single-document summarization for news) and showed an empirical distribution of summary length over document size. However, the length problem has been gradually ignored later, since researchers need to fix the length so as to estimate different summarization models conveniently. A typical instance is the Document Understanding Conferences (DUC)[1], which provide authoritative evaluation for summarization systems. The DUC conferences collect news aritcles as the input data and define various summarization tasks, such as generic multi-document summarization, query-focused summarization and update summarization. In all the DUC tasks, the output is restricted within a length. Then human-generated

[1] After 2007, the DUC tasks are incorporated into the Text Analysis Conference (TAC).

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 736–746,
Seattle, Washington, USA, 18-21 October 2013.

summaries are provided to evaluate the results of different summarization systems. Limiting the length of summaries contributed a lot to the development of summarization techniques, but as we discussed before, in many cases keeping the summaries of the same size is not a good choice.

Moreover, even in constant-length summarization, how to define a proper size of summaries for the summarization tasks is quite a problem. Why does DUC2007 main task require 250 words while Update task require 100 words? Is it reasonable? A short summary may sacrifice the coverage, while a long summary may cause redundance. Automatically determining the best size of summaries according to the input documents is valuable, and it may deepen our understanding of summarization.

In this work, we aim to find the proper length for document summarization automatically and generate varying-length summaries based on the document itself. The varying-length summarization is more robust for unbalanced clusters. It can also provide a recommended size as the predefined summary length for general constant-length summarization systems. We advance a Bayesian nonparametric model of extractive multi-document summarization to achieve this goal. As far as we are concerned, it is the first model that can learn appropriate lengths of summaries.

Bayesian nonparametric (BNP) methods are powerful tools to determine the size of latent variables (Gershman and Blei, 2011). They let the data ”speak for itself” and allow the dimension of latent variables to grow with the data. In order to integrate the BNP methods into document summarization, we follow the assumption that the original documents should be recovered from the reconstruction of summaries (Ma and Wan, 2010; He et al., 2012). We use the Beta process as a prior to generate binary vectors for selecting active sentences that reconstruct the original documents. Then we construct a Bayesian framework for summarization and use the variational approximation for inference. Experimental results on DUC2004 dataset demonstrate the effectiveness of our model. Besides, we reorganize the original documents to generate some new datasets, and examine how the summary length changes on the new data. The results prove that our summary length determination is rational and necessary on unbalanced data.

2 Related Work

2.1 Research on Summary Length

Summary length is an important aspect for generating and evaluating summaries. Early research on summary length (Goldstein et al., 1999) focused on discovering the properties of human-generated summaries and analyzing the effect of compression ratio. It demonstrated that an evaluation of summarization systems must take into account both the compression ratios and the characteristics of the documents. Radev and Fan (2000) compared the readability and speedup in reading time of 10% summaries and 20% summaries[2] for topic sets with different number of documents. Sweeney et al. (2008) developed an incremental summary containing additional sentences that provide context. Kaisser et al. (2008) studied the impact of query types on summary length of search results. Other than the content of original documents, there are also some other factors affecting summary length especially in specific applications. For example, Sweeney and Crestani (2006) studied the relation between screen size and summary length on mobile platforms. The conclusion of their work is the optimal summary size always falls into the shorter one regardless of the screen size.

In sum, the previous works on summary length mostly put their attention on the empirical study of the phenomenon, factors and impacts of summary length. None of them automatically find the best length, which is our main task in this paper. Nevertheless, they demonstrated the importance of summary length in summarization and the reasonability of determining summary length based on content of news documents (Goldstein et al., 1999) or search results (Kaisser et al., 2008). As our model is mainly applied for generic summarization of news articles, we do not consider the factor of screen size in mobile applications.

2.2 BNP Methods in Document Summarization

Bayesian nonparametric methods provide a Bayesian framework for model selection and adaptation using nonparametric models (Gershman

[2] 10% and 20% are the compression rates, and the documents are from search results in information retrieval systems.

and Blei, 2011). A BNP model uses an infinite-dimensional parameter space, but invokes only a finite subset of the available parameters on any given finite data set. This subset generally grows with the data set. Thus BNP models address the problem of choosing the number of mixture components or latent factors. For example, the hierarchical Dirichlet process (HDP) can be used to infer the number of topics in topic models or the number of states in the infinite Hidden Markov model (Teh et al., 2006).

Recently, some BNP models are also involved in document summarization approaches (Celikyilmaz and Hakkani-Tür, 2010; Chang et al., 2011; Darling and Song, 2011). BNP priors such as the nested Chinese restaurant process (nCRP) are associated with topic analysis in these models. Then the topic distributions are used to get the sentence scores and rank sentences. BNP here only impacts the number and the structure of the latent topics, but the summarization framework is still constant-length. Our BNP summarization model differs from the previous models. Besides using the HDP for topic analysis, our approach further integrates the beta process into sentence selection. The BNP method in our model are directly used to determine the number of summary sentences but not latent topics.

3 BNP Summarization

In this section, we first introduce the BNP priors which will be used in our model. Then we propose our model called BNP summarization.

3.1 The Beta Process and the Bernoulli process

The beta process(BP) (Thibaux and Jordan, 2007; Paisley and Carin, 2009) and the related Indian buffet process(IBP) (Griffiths and Ghahramani, 2005) are widely applied to factor/feature analysis. By defining the infinite dimensional priors, these factor analysis models need not to specify the number of latent factors but automatically determine it.

Definition of BP (Paisley et al., 2010): Let B_0 be a continuous measure on a space Θ and $B_0(\Theta) = \gamma$. If B_k is defined as follows,

$$\begin{aligned} B_k &= \sum_{k=1}^{N} \pi_k \delta_{\theta_k}, \\ \pi_k &\sim Beta(\frac{\alpha\gamma}{N}, \alpha(1-\frac{\gamma}{N})) \\ \theta_k &\sim \frac{1}{\gamma} B_0 \end{aligned} \quad (1)$$

(where δ_{θ_k} is the atom at the location θ_k; and α is a positive scalar), then as $N \to \infty, B_k \to B$ and B is a beta process: $B \sim BP(\alpha B_0)$.

Finite Approximation: The beta process is defined on an infinite parameter space, but sometimes we can also use its finite approximation by simply setting N to a large number (Paisley and Carin, 2009).

Bernoulli Process: The beta process is conjugate to a class of Bernoulli processes, denoted by $X \sim Bep(B)$. If B is discrete, of the form in (1), then $X = \sum_k b_k \delta_{\theta_k}$ where the b_k are independent Bernoulli variables with the probability $p(b_k = 1) = \pi_k$. Due to the conjugation between the beta process priors and Bernoulli process, the posterior of B given M samples $X_1, X_2, ...X_M$ where $X_i \sim Bep(B) for i = 1,,, M.$ is also a beta process which has updated parameters:

$$\begin{aligned} &B|X_1, X_2, ..., X_M \\ &\sim BP(\alpha + M, \tfrac{\alpha}{\alpha+M} B_0 + \tfrac{1}{c+M} \textstyle\sum_i X_i) \end{aligned} \quad (2)$$

Application of BP: Furthermore, marginalizing over the beta process measure B and taking $\alpha = 1$, provides a predictive distribution on indicators known as the Indian buffet process (IBP) (Thibaux and Jordan, 2007). The beta process or the IBP is often used in a feature analysis model to generate infinite vectors of binary indicator variables(Paisley and Carin, 2009), which indicates whether a feature is used to represent a sample. In this paper, we use the beta process as the prior to select sentences.

3.2 Framework of BNP Summarization

Most existing approaches for generic extractive summarization are based on sentence ranking. However, these methods suffer from a severe problem that they cannot make a good trade-off between the coverage and minimum redundancy (He et al., 2012). Some global optimization algorithms are developed, instead of greedy search, to select the best overall summaries (Nenkova and McKeown, 2012). One approach to global optimization of summarization is to regard the summarization as a reconstruction process (Ma and Wan, 2010; He et al., 2012) . Considering a good summary must catch most of the important information in original documents, the original documents are assumed able to be recovered from summaries with some information loss. Then the summarization problem is turned into finding the sentences that cause the least reconstruction error (or information loss). In this paper, we follow the assumption and formulate summarization as a Bayesian framework.

First we review the models of (Ma and Wan, 2010) and (He et al., 2012). Given a cluster of M documents $x_1, x_2, ..., x_M$ and the sentence set contained in the documents as $S = [s_1, s_2, ..., s_N]$, we denote all corresponding summary sentences as $V = [v_1, ..., v_n]$, where n is the number of summary sentences and N is the number of all sentences in the cluster. A document x_i and a sentence v_i or s_i here are all represented by weighted term frequency vectors in the space $\mathbb{R}^d$, where d is the number of total terms (words).

Following the reconstruction assumption, a candidate sentence v_i can be approximated by the linear combination of summary sentences: $s_i \simeq \sum_{j=1}^{n} w'_j v_j$, where w'_j is the weight for summary sentence v_j. Thus the document can also be approximately represented by a linear combination of summary sentences (because it is the sum of the sentences).

$$x_i \simeq \sum_{j=1}^{n} w_j v_j. \tag{3}$$

Then the work in (He et al., 2012) aims to find the summary sentence set that can minimize the reconstruction error $\sum_{i=1}^{N} ||s_i - \sum_{j=1}^{n} w'_j v_j||^2$; while the work in (Ma and Wan, 2010) defines the problem as finding the sentences that minimize the distortion between documents and its reconstruction $dis(x_i, \sum_{j=1}^{n} w_j v_j)$ where this distortion function can also be a squared error function.

Now we consider the reconstruction for each document, if we see the document x_i as the dependent variable, and the summary sentence set S as the independent variable, the problem to minimize the reconstruction error can be seen as a linear regression model. The model can be easily changed to a Bayesian regression model by adding a zero-mean Gaussian noise ϵ (Bishop, 2006), as follows.

$$x_i = \sum_{j=1}^{n} w_j v_j + \epsilon_i \tag{4}$$

where the weights w_j are also assigned a Gaussian prior.

The next step is sentence selection. As our system is an extractive summarization model, all the summary sentences are from the original document cluster. So we can use a binary vector $z_i =< z_{i1}, ..., z_{iN} >^T$ to choose the active sentences V (i.e. summary sentences) from the original sentence set S. The Equation (4) is turned into $x_i = \sum_{j=1}^{N} \phi_{ij} * z_{ij} s_j + \epsilon_i$. Using a beta process as a prior for the binary vector z_i, we can automatically infer the number of active component associated with z_i. As to the weights of the sentences, we use a random vector ϕ_i which has the multivariate normal distribution because of the conjugacy. $\phi_i \in \mathbb{R}^N$ is an extension to the weights $\{w_1, ...w_n\}$ in (4).

Integrating the *linear reconstruction* (4) and the *beta process*[3] (1), we get the complete process of summary sentence selection as follows.

$$\begin{aligned} x_i &= S(\phi_i \circ z_i) + \epsilon_i \\ S &= [s_1, s_2, ..., s_N] \\ z_{ij} &\sim Bernoulli(\pi_j) \\ \pi_j &\sim Beta(\frac{\alpha\gamma}{N}, \alpha(1 - \frac{\gamma}{N})) \\ \phi_i &\sim \mathcal{N}(0, \sigma_\phi^2 I) \\ \epsilon_i &\sim \mathcal{N}(0, \sigma_\epsilon^2 I) \end{aligned} \tag{5}$$

where N is the number of sentences in the whole document cluster. The symbol $\circ$ represents the elementwise multiplication of two vectors.

One problem of the reconstruction model is that the word vector representation of the sentences are sparse, which dramatically increase the reconstruction error. So we bring in topic models to reduce the

[3] We use the finite approximation because the number of sentences is large but finite

dimension of the data. We use a HDP-LDA (Teh et al., 2006) to get topic distributions for each sentence, and we represent the sentences and documents as the topic weight vectors instead of word weight vectors. Finally x_i is a K-dimensional vector and S is a $K * N$ matrix, where K is the number of topics in topic models.

4 Variational Inference

In this section, we derive a variational Bayesian algorithm for fast inference of our sentence selection model. Variational inference (Bishop, 2006) is a framework for approximating the true posterior with the best from a set of distributions $Q : q* = \arg\min_{q \in Q} KL(q(Z)|p(Z|X))$. Suppose $q(Z)$ can be partitioned into disjoint groups denoted by Z_j, and the q distribution factorizes with respect to these groups: $q(Z) = \prod_{j=1}^{M} q(Z_j)$. We can obtain a general expression for the optimal solution $q_j^*(Z_j)$ given by

$$\ln q_j^*(Z_j) = \mathbb{E}_{i \neq j}[\ln p(X, Z)] + const. \quad (6)$$

where $\mathbb{E}_{i \neq j}[\ln p(X, Z)]$ is the expectation of the logarithm of the joint probability of the data and latent variables, taken over all variables not in the partition. We will therefore seek a consistent solution by first initializing all of the factors $q_j(Z_j)$ appropriately and then cycling through the factors and replacing each in turn with a revised estimate given by (6) evaluated using the current estimates for all of the other factors.

Update for Z

$$p(z_{ij}|\pi_j, x_i, S, \phi_i) \propto p(x_i|z_{ij}, s_j, \phi_i)p(z_{ij}|\pi_j)$$

We use $q(z_{ij})$ to approximate the posterior:

$$\begin{aligned}
&q(z_{ij}) \\
&\propto \exp\{\mathrm{E}[\ln(p(x_i|z_{ij}, z_i^{-j}, S, \phi_i)) + \ln(p(z_{ij}|\pi))]\} \\
&\propto \exp\{\mathrm{E}[\ln(\pi_j)]\}* \\
&\exp\{\mathrm{E}[-\frac{1}{2\sigma_\epsilon^2}\left(x_i^{-j} - s_j z_{ij}\phi_{ij}\right)^T \left(x_i^{-j} - s_j z_{ij}\phi_{ij}\right)]\} \\
&\propto \exp\{\overline{\ln(\pi_j)}\}* \\
&\exp\{-\frac{\left(\overline{\phi_{ij}^2} * \overline{z_{ij}^2} * \overline{s_j^T s_j} - 2\overline{\phi_{ij}} * \overline{z_{ij}} * \overline{s_j}^T * \overline{x_i^{-j}}\right)}{2\sigma_\epsilon^2}\}
\end{aligned} \quad (7)$$

where $x_i^{-j} = x_i - S^{-j}(\phi_i^{-j} \circ z_i^{-j})$, and the symbol $\bar{}$ indicates the expectation value. The $\overline{\phi_{ij}^2}$ can be extended to this form:

$$\overline{\phi_{ij}^2} = \overline{\phi_{ij}}^2 + \Delta_i^j \quad (8)$$

where Δ_i^j means the j^{th} diagonal element of Δ_i which is defined by Equation 13.

As z_i is a binary vector, we only calculate the probability of $z_{ij} = 1$ and $z_{ij} = 0$.

$$\begin{aligned}
&q(z_{ij} = 1) \propto \exp\{\overline{\ln(\pi_j)}\} * \\
&\exp\{-\frac{1}{2\sigma_\epsilon^2}\left(\overline{\phi_{ij}^2} * \overline{s_j^T s_j} - 2\overline{\phi_{ij}} * \overline{s_j}^T * \overline{x_i^{-j}}\right)\} \\
&q(z_{ij} = 0) \propto \exp\{\overline{\ln(1 - \pi_j)}\}
\end{aligned} \quad (9)$$

The expectations can be calculated as

$$\overline{\ln(\pi_j)} = \varphi(\frac{\alpha\gamma}{N} + \overline{n_j}) - \varphi(\alpha + M) \quad (10)$$

$$\overline{\ln(1 - \pi_j)} = \varphi(\alpha(1 - \frac{\gamma}{N}) + M - \overline{n_j}) - \varphi(\alpha + M) \quad (11)$$

where $\overline{n_j} = \sum_{i=1}^{M} z_{ij}$.

Update for π

$$p(\pi_j|Z) \propto p(\pi_j|\alpha, \gamma, N)p(Z|\pi_j)$$

Because of the conjugacy of the beta to Bernoulli distribution, the posterior of π is still a beta distribution:

$$\pi_j \sim Beta(\frac{\alpha\gamma}{N} + \overline{n_j}, \alpha(1 - \frac{\gamma}{N}) + M - \overline{n_j}) \quad (12)$$

Update for Φ

$$p(\phi_i|x_i, Z, S) \propto p(x_i|\phi_i, Z, S)p(\phi_i|\sigma_\phi^2)$$

The posterior is also a normal distribution with mean μ_i and covariance Δ_i.

$$\Delta_i = \left(\frac{1}{\sigma_\epsilon^2}\overline{\tilde{S}_i^T \tilde{S}_i} + \frac{1}{\sigma_\phi^2}I\right)^{-1} \quad (13)$$

$$\mu_i = \Delta_i \left(\frac{1}{\sigma_\epsilon^2}\overline{\tilde{S}_i}^T x_i\right) \quad (14)$$

Here $\tilde{S}_i \equiv S \circ \tilde{z}_i$ and $\tilde{z}_i \equiv [z_i, ..., z_i]^T$ is a $K \times N$ matrix with the vector z_i repeated K(the number of the latent topics) times.

$$\overline{\tilde{S}_i} = S * \overline{\tilde{z}_i} \quad (15)$$

$$\overline{\tilde{S}_i^T \tilde{S}_i} = (S^T S) \circ (\overline{z_i} * \overline{z_i}^T + Bcov_i) \tag{16}$$

$$Bcov_i = \text{diag}[\overline{z_{i1}}(1-\overline{z_{i1}}), ..., \overline{z_{iN}}(1-\overline{z_{iN}})] \tag{17}$$

Update for σ_ϵ^2

$$p(\sigma_\epsilon^2 | \Phi, X, Z, S) \propto p(X | \Phi, Z, S, \sigma_\epsilon^2) p(\sigma_\epsilon^2)$$

By using a conjugate prior, inverse gamma prior $InvGamma(u, v)$, the posterior can be calculated as a new inverse gamma distribution with parameters

$$\begin{aligned} u' &= u + MK/2 \\ v' &= v + \frac{1}{2}\sum_{i=1}^{M}(||x_i - S(\overline{z_i} \circ \overline{\phi_i})|| + \xi_i) \end{aligned} \tag{18}$$

where

$$\begin{aligned} \xi_i = \textstyle\sum_{j=1}^{N} (\overline{z_{ij}^2} * \overline{\phi_{ij}^2} * s_j^T s_j - \overline{z_{ij}}^2 * \overline{\phi_{ij}}^2 * s_j^T s_j) \\ + \textstyle\sum_{j \neq l} \overline{z_{ij}} * \overline{z_{il}} * \Delta_{i,jl} * s_j^T s_l \end{aligned}$$

Update for σ_ϕ^2

$$p(\sigma_\phi^2 | \Phi) \propto p(\Phi | \sigma_\phi^2) p(\sigma_\phi^2)$$

By using a conjugate prior, inverse gamma prior $InvGamma(e, f)$, the posterior can be calculated as a new inverse gamma distribution with parameters

$$\begin{aligned} e' &= e + MN/2 \\ f' &= f + \frac{1}{2}\sum_{i=1}^{M}\left((\overline{\Phi})^T \overline{\Phi} + trace(\Delta_i')\right) \end{aligned} \tag{19}$$

5 Experiments

To test the capability of our BNP summarization systems, we design a series of experiments. The aim of the experiments mainly includes three aspects:

1. To demonstrate the summaries extracted by our model have good qualities and the summary length determined by the model is reasonable.
2. To give examples where varying summary length is necessary.
3. To observe the distribution of summary length.

We evaluate the performance on the dataset of DUC2004 task2. The data contains 50 document clusters, with 10 news articles in each cluster. Besides, we construct three new datasets from the DUC2004 dataset to further prove the advantage of variable-length summarization. We separate each cluster in the original dataset into two parts where each has 5 documents, hence getting the *Separate* Dataset; Then we randomly combine two original clusters in the DUC2004 dataset, and get two datasets called *Combined1* and *Combined2*. Thus each of the clusters in the combined datasets include 20 documents with two different themes.

5.1 Evaluation of Summary Qualities

First, we implement our BNP summarization model on the DUC2004 dataset, with summary length not limited. At the topic analysis step, we use the HDP model and follow the inference in (Teh et al., 2006). For the sentence selection step, we use the variational inference described in Section 4, where the parameters in the beta process (5) are set as $\gamma = 1, \alpha = 1$. The summaries that we finally generate have an average length of 164 words. We design several popular unsupervised summarization systems and compare them with our model.

- The *Random* model selects sentences randomly for each document cluster.
- The *MMR* (Carbonell and Goldstein, 1998) strives to reduce redundancy while maintaining relevance. For generic summarization, we replace the query relevance with the relevance to documents.
- The *Lexrank* model (Erkan and Radev, 2004) is a graph-based method which choose sentences based on the concept of eigenvector centrality.
- The *Linear* Representation model (Ma and Wan, 2010) has the same assumption as ours and it can be seen as an approximation of the constant-length version of our model.

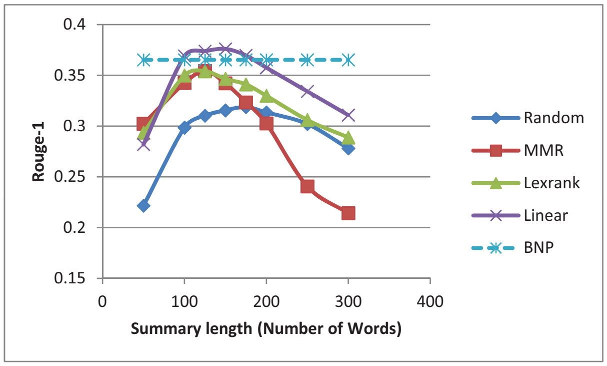

Figure 1: Rouge-1 values on DUC2004 dataset.

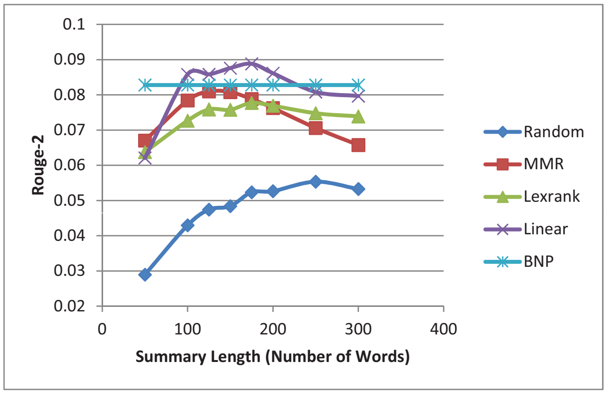

Figure 2: Rouge-2 values on DUC2004 dataset.

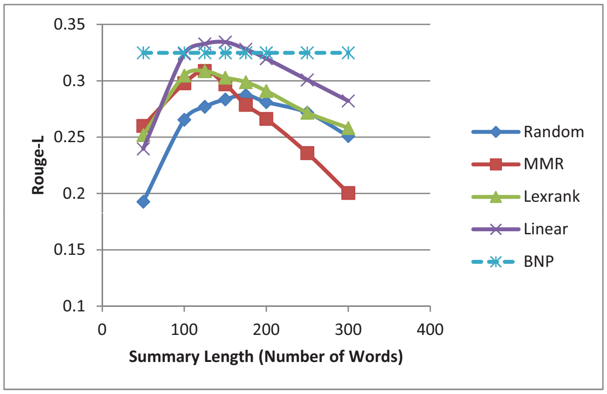

Figure 3: Rouge-L values on DUC2004 dataset.

All the compared systems are implemented at different predefined lengths from 50 to 300 words. Then we evaluate the summaries with ROUGE[4] tools (Lin and Hovy, 2003) in terms of the f-measure scores of Rouge-1 Rouge-2, and Rouge-L. The metric of Rouge f-measure takes into consideration the summary length in evaluation, so it is proper for our experiments. From Fig.1, Fig.2 and Fig.3, we can see that the result of BNP summarization (the dashed line) gets the second best value among all systems. It is only defeated by the *Linear* model but the result is comparable to the best in Fig.1 and Fig.3; while it exceeds other systems at all lengths. This proves the good qualities of our BNP summaries. The reason that the *Linear* system gets a little better result may be its weights for linear combination of summary sentences are guaranteed nonnegative while in our model the weights are zero-mean Gaussian variables. This may lead to less redundance in sentence selection for the *Linear* Representation model.

Turn to the length determination. We take advantage of the *Linear* Representation model to approximate the constant-length version of our model. Comparing the summaries generated at different predefined lengths, Fig.4 shows the the model gets the best performance (Rouge values) at the length around 164 words, the length learned by our BNP model. This result partly demonstrates our length determination is rational and it can be used as the recommended length for some constant-length summarization systems, such as the *Linear* .

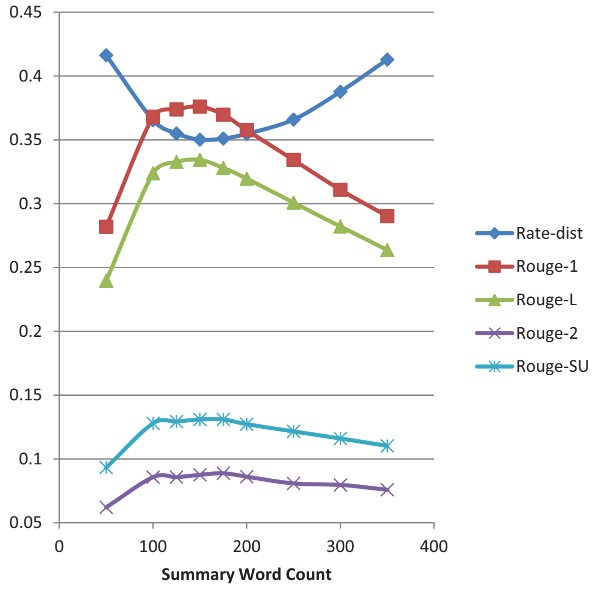

Figure 4: Rate-dist value V.S. summary word length.

[4] we use ROUGE1.5.5 in this work.

5.2 A New Evaluation Metric

The Rouge evaluation requires golden standard summaries as the base. However, in many cases we cannot get the reference summaries. For example, when we implement experiments on our expanded datasets (the separate and combined clusters of documents), we do not have exact reference summaries. Louis and Nenkova (2009) advanced an automatic summary evaluation without human models. They used the Jensen-Shannon divergence(JSD) between the input documents and the summaries as a feature, and got high correlation with human evaluations and the rouge metric. Unfortunately, it was designed for comparison at a constant-length, which cannot meet our needs. To extend the JSD evaluation to compare varying-length summaries, we propose a new measure based on information theory, the rate-distortion (Cover and Thomas, 2006).

Rate-Distortion: The distortion function $d(x, \hat{x})$ is a measure of the cost of representing the symbol x to a new symbol $\hat{x}$; and the rate can indicate how much compression can be achieved. The problem of finding the minimum rate can be solved by minimizing the functional

$$\mathcal{F}[p(\hat{x}|x)] = I(X; \hat{X}) + \beta\mathbb{E}(d(x, \hat{x})). \qquad (20)$$

where $I(X; \hat{X})$ denotes the mutual information. The rate-distortion theory is a fundamental theory for lossy data compression. Recently, it has also been successfully employed for text clustering (Slonim, 2002) and document summarization (Ma and Wan, 2010). Slonim (2002) claims that the mutual information $I(X; \hat{X})$ measures the compactness of the new representation. Thus the rate-distortion function is a trade-off between the compactness of new representation and the expected distortion. Specifically in summarization, the summaries can be seen as the new representation $\hat{X}$ of original documents X. A good summary balances the compression ratio and the information loss, thus minimizing the function (20). So we use the function (20)(we set $\beta = 1$) to compare which summary is a better compression. The JS-divergence (JSD), which has been proved to have high correlation with manual evaluation (Louis and Nenkova, 2009) for constant-length summary evaluation, is utilized as the distortion in the function. In the following sections, we simply call the values of the function (20) *rate-dist*. In fact, the rate-dist values can be seen as the JSD measure with length regularization.

To check the effectiveness of rate-dist measure, we evaluate all summaries generated in Section 5.1 with the new measure (the lower the better). Fig. 5 shows that the results accord with the ones in Fig. 1 and Fig. 3. Moreover, in Fig. 4, the curve of rate-dist values has a inverse tendency of Rouge measures (Rouge-1, Rouge-2, Rouge-L and Rouge-SU4 are all listed here), and the best performance also occurs around the summary length of 164 words. This even more clearly reveals that the BNP summarization achieves a perfect tradeoff between compactness and informativeness. Due to the accordance with rouge measures, it is promising to be regarded as an alternative to the rouge measures in case we do not have reference summaries.

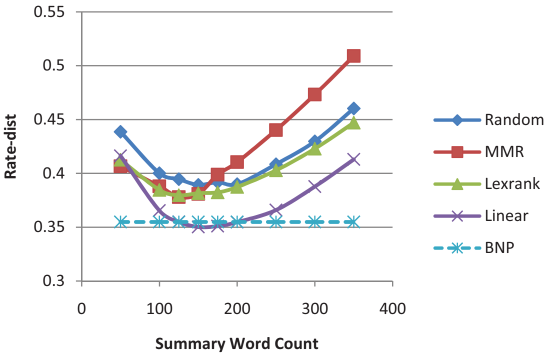

Figure 5: Comparison of BNP Summarization with other systems using rate-dist measure.

5.3 Necessity of Varying Summary Length

In this section, we discuss the necessity of length determination and how summary length changes according to the input data. As explained before, we generate three new datasets from the original DUC2004 dataset. Now we use them to indicate varying summary length is necessary when the input data varies a lot.

Table 1 shows the average summary length of different data sets. The results satisfy the intuitive expectation of summary length change. When we split a 10-document cluster into two 5-document parts, we expect the average summary length of the new clusters to be a little smaller than the original cluster but much larger than half of the original length, because all the documents concentrate on the same themes. When we combine two clusters into one, the summary length should be smaller than the sum of the summary lengths of two original clusters due to some unavoidable common background information but much larger than the summary length of original clusters.

Original	Separate	Combined1	Combined2
164	115	250	231

Table 1: Average summary length (number of words) on different datasets

We also run the Linear Representation system at different lengths on the new datasets and evaluate the qualities. As we do not have golden standard for the new datasets, so we only use the rate-dist measure here. Results in Table 2,3,4 show the summaries which do not change the predefined length [5] perform significantly worse than the BNP summarization. All the comparison is statistically significant. So varying summary length is necessary when the input changes a lot, and our model can just give a good match to the new data. This characteristic also can be used to give recommended summary length for extractive summarization systems when given unknown data.

	Predefined	Unchanged	BNP
Length	665 bytes	164 words	115 words
Rate-dist	0.4130	0.4404	0.4007

Table 2: Comparison of summary lengths on Separate Dataset.

	Predefined	Unchanged	BNP
Length	665 bytes	164 words	250 words
Rate-dist	0.3768	0.3450	0.3238

Table 3: Comparison of summary lengths on Combined1 Dataset.

Then we observe the summary length distributions and compression ratios according to document size(the length of the whole documents in a cluster). The average summary length increases (Fig. 6), while the compression ratios decreases (Fig. 7) as document size grows. The rule of the compression ratio here agrees with the rule in (Goldstein et al., 1999), although that work is done for single-document summarization.

	Predefined	Unchanged	BNP
Length	665 bytes	164 words	231 words
Rate-dist	0.3739	0.3464	0.3326

Table 4: Comparison of summary lengths on Combined2 Dataset.

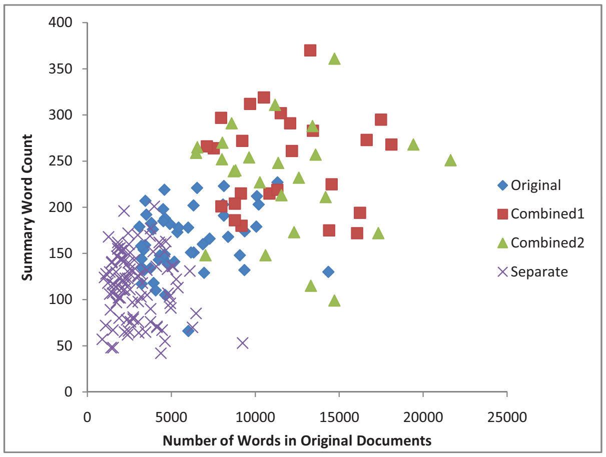

Figure 6: The distribution of summary word length.

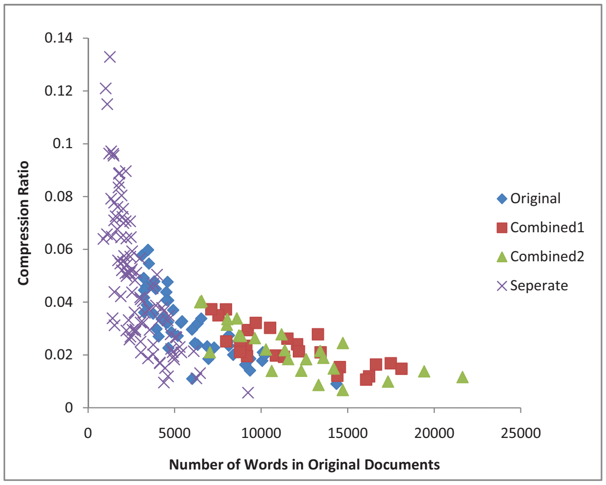

Figure 7: Compression ratio versus document word length.

[5] 665 bytes is the DUC2004 requirement and 164 words is the best length on original data

6 Conclusion and Future Work

In this paper, we present a new problem of finding a proper summary length for multi-document summarization based on the document content. A Bayesian nonparametric model is proposed to solve this problem. We use the beta process as the prior to construct a Bayesian framework for summary sentence selection. Experimental results are shown on DUC2004 dataset, as well as some expanded datasets. We demonstrate the summaries we extract have good qualities and the length determination of our system is rational.

However, there is still much work to do for variable-length summarization. First, Our system is extractive-base summarization, which cannot achieve the perfect coherence and readability. A system which can determine the best length even for abstractive summarization will be better. Moreover, in this work we only consider the aspect of data compression and evaluate the performance using an information-theoretic measure. In future we may consider more human factors, and prove the summary length determined by our system agrees with human preference. In addition, in the experiments, we only use the imbalanced datasets as the example that intuitively needs varying the summary length. However, the data type is also important to impact the summary length. In future, we may extend the work by studying more cases that need varying summary length.

References

Christopher M. Bishop. 2006. *Pattern recognition and machine learning*. . Vol. 4. No. 4. New York: springer.

Jaime Carbonell, and Jade Goldstein. 1998. The Use Of Mmr, Diversity-Based Reranking For Reordering Documents And Producing Summaries. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1998.*

Asli Celikyilmaz and Dilek Hakkani-Tür. 2010. A Hybrid Hierarchical Model for Multi-Document Summarization. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 815-824.

Ying-Lan Chang, Jui-Jung Hung and Jen-Tzung Chien 2011. Bayesian Nonparametric Modeling Of Hierarchical Topics And Sentences. *IEEE International Workshop on Machine Learning for Signal Processing*, September 18-21, 2011, Beijing, China.

Thomas M. Cover, and Joy A. Thomas. 2006. *Elements of information theory*. Wiley-interscience, 2006.

William M. Darling and Fei Song. 2011. PathSum: A Summarization Framework Based on Hierarchical Topics. *Canadian AI Workshop on Text Summarization*, St. John's, Newfoundland.

Samuel J. Gershman and David M. Blei. 2011. A Tutorial On Bayesian Nonparametric Models. *Journal of Mathematical Psychology(2011).*

Thomas L. Griffiths and Zoubin Ghahramani. 2005. Infinite Latent Feature Models and the Indian Buffet Process. *Advances in Neural Information Processing Systems 18.*

Jade Goldstein, Mark Kantrowitz, Vibhu Mittal and Jaime Carbonelly. 1999. Summarizing Text Documents: Sentence Selection and Evaluation Metrics. *Proceedings of SIGIR'99* , pages 121-128.

Zhanying He, Chun Chen, Jiajun Bu, CanWang, Lijun Zhang, Deng Cai and Xiaofei He. 2012. Document Summarization Based on Data Reconstruction. *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Michael Kaisser, Marti A. Hearst, John B. Lowe. 2008. Improving Search Results Quality by Customizing Summary Lengths. *Proceedings of ACL-08: HLT*, pages 701-709.

Chin-Yew Lin, Guihong Cao, Jianfeng Gao, and Jian-Yun Nie. 2006. An Information-Theoretic Approach to Automatic Evaluation of Summaries. *Proceedings of NAACL2006*, pages 463-470.

Chin-Yew Lin, and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. *Proceedings of NAACL2003.*

Annie Louis and Ani Nenkova. 2009. Automatically Evaluating Content Selection in Summarization without Human Models. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 306-314. Singapore, 6-7 August 2009.

Tengfei Ma and Xiaojun Wan. 2010. Multi-document Summarization Using Minimum Distortion. *IEEE 10th International Conference on Data Mining (ICDM).*

Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. *Mining Text Data, Chapter 3, Springer Science+Business Media, LLC* (2012).

John Paisley and Lawrence Carin. 2009. Nonparametric Factor Analysis with Beta Process Priors. *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada.

John Paisley, Aimee Zaas, Christopher W. Woods, Geoffrey S. Ginsburg and Lawrence Carin. 2010. A Stick-Breaking Construction of the Beta Process. *Proceedings of the 27 th International Confer- ence on Machine Learning*, Haifa, Israel, 2010.

Dragomir R. Radev and Weiguo Fan. 2000. Effective search results summary size and device screen size: Is there a relationship. *Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval*

Günes Erkan, and Dragomir R. Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22 (2004) 457-479.

Noam Slonim. 2002. The Information Bottleneck: Theory and Applications. *PHD Thesis of the Hebrew University* .

Simon Sweeney and Fabio Crestani. 2006. Effective search results summary size and device screen size: Is there a relationship. *Information Processing and Management* 42 (2006) 1056-1074.

Simon Sweeney, Fabio Crestani and David E. Losada. 2008. 'Show me more': Incremental length summarisation using novelty detection. *Information Processing and Management* 44 (2008) 663-686.

Yee Whye Teh, Dilan Görür, and Zoubin Ghahramani. 2007. Stick-breaking Construction for the Indian Buffet Process. *Proceedings of the International Conference on Artificial Intelligence and Statistics*.

Y.W. Teh, M.I. Jordan, M.J. Beal and D.M. Blei. 2006. Hierarchical Dirichlet Processes. *JASA* , 101(476):1566-1581.

Romain Thibaux and Michael I. Jordan. 2009. Hierarchical Beta Processes and the Indian Buffet Process. *AISTATS2007*.