An Information-Theoretic Approach to Automatic Evaluation of Summaries

Chin-Yew Lin[+], Guihong Cao[*], Jianfeng Gao[#], and Jian-Yun Nie[*]

Information Sciences Institute[+]
University of Southern California
Marina del Rey, CA 90292
USA
cyl@isi.edu

Université de Montréal[*]
Montréal, Canada
{caogui,nie}@iro.umontreal.ca

Microsoft Corporation[#]
One Microsoft Way
Redmond, WA 98052
USA
jfgao@microsoft.com

Abstract

Until recently there are no common, convenient, and repeatable evaluation methods that could be easily applied to support fast turn-around development of automatic text summarization systems. In this paper, we introduce an information-theoretic approach to automatic evaluation of summaries based on the Jensen-Shannon divergence of distributions between an automatic summary and a set of reference summaries. Several variants of the approach are also considered and compared. The results indicate that JS divergence-based evaluation method achieves comparable performance with the common automatic evaluation method ROUGE in single documents summarization task; while achieves better performance than ROUGE in multiple document summarization task.

1 Introduction

Most previous automatic evaluation methods in summarization use co-occurrence statistics (Lin and Hovy 2003) to measure the content overlap between an automatic summary and a set of reference summaries. Among them, ROUGE (Lin 2004) has been used by the annual summarization evaluation conference, Document Understanding Conference[1] (DUC), sponsored by NIST since 2001. Content and quality of a summary are the two main aspects of summarization measured in the past DUCs. Using a manual evaluation interface called SEE[2], NIST assessors compared the content overlap between a system summary and a reference summary and assigned a coverage score to indicate the extent of the overlap between system and reference summaries. The overall system content coverage score was then the average of coverage scores over a set of test topics. NIST assessors also judged the quality of a peer summary by answering a set of quality assessment questions related to grammaticality, coherence, and organization for each system summary. However, we only focus on automatic evaluation of content coverage in this paper and aim at establishing a statistical framework that can perform at least as good as the current state-of-the-art automatic summarization evaluation methods such as ROUGE.

We start with a brief description of our statistical summary generation model and how to estimate its parameters in the next section. We then describe experimental setups and criterion of success in Section 3. The results of the experiments are shown and analyzed in Section 4. We discuss related work and recent advances in statistical language models for information retrieval in Section 5. Finally, we conclude and suggest future directions in Section 6.

2 Summarization Evaluation Using Information-Theoretic Measures

Given a set of documents $D = \{d_1, d_2, \ldots, d_i\}$[3], $i = 1$ to n, we assume there exists a probabilistic distribution with parameters specified by θ_R that generates reference summaries from D. The task of summarization is to estimate θ_R. Similarly, we as-

[1] Please see: http://duc.nist.gov for more information.

[2] SEE can be downloaded at: http://www.isi.edu/~cyl/SEE.

[3] $n = 1$ for a single document summarization task; $n > 1$ for a multi-document summarization task.

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 463–470,
New York, June 2006.

sume every system summary is generated from a probabilistic distribution with parameters specified by θ_A. Therefore, a good summarizer should have its θ_A very close to θ_R and the process of summary evaluation could be viewed as a task of estimating the distance between θ_A and θ_R.

For example, if we use Kullback-Leibler (*KL*) divergence as the distance function, then the summary evaluation task could be viewed as finding the *KL* divergence between θ_A and θ_R. However, *KL* divergence is unspecified when a value is defined in θ_R but not in θ_A (Lin 1991, Dagan et al. 1999). Usually smoothing has to be applied to address this missing data problem (unseen word in this case).

Another problem is that *KL* divergence is not symmetric, i.e. $KL(\theta_R \| \theta_A) \neq KL((\theta_A \| \theta_R)$, except when $\theta_R = \theta_A$. This is counter-intuitive in our application scenario. We therefore use generalized Jensen-Shannon (*JS*) divergence proposed by Lin (1991). The *JS* divergence can be written as follows:

$$JS_\pi(p_1, p_2, \cdots, p_n) = H\left(\sum_{i=1}^{n} \pi_i p_i\right) - \sum_{i=1}^{n} \pi_i H(p_i), \quad (1)$$

where p_i is a probability distribution with weight π_i, $\sum_{i=1}^{n} \pi_i = 1$, and $H(\cdot)$ is Shannon entropy. For $n = 2$ and equal weight, we have the following:

$$JS_{1/2}(p_1 \| p_2) = \frac{1}{2} \sum_{\theta_i} \left(p_1 \log\left(\frac{p_1}{\frac{1}{2}p_1 + \frac{1}{2}p_2} \right) + p_2 \log\left(\frac{p_2}{\frac{1}{2}p_1 + \frac{1}{2}p_2} \right) \right) \quad (2)$$

Since the Jensen-Shannon divergence is a distance measure, we take its negative value to indicate the similarity between two distributions as follows:

$$Score^{JS}_{summary}(S_A \mid S_R) = -JS_{1/2}(p(\theta_A \mid S_A) \| p(\theta_R \mid S_R)). \quad (3)$$

Equation (3) suggests that the problem of summary evaluation could be cast as ranking system summaries according to their negative Jensen-Shannon divergence with respect to the estimated posterior distribution of reference summaries. The question now is how to estimate these distributions.

2.1 Estimation of Posterior and Prior System Summary Distributions

θ_A is estimated via maximum *a posterior* (MAP) as:

$$\theta_A^{MP} = \arg\max_{\theta_A} p(\theta_A \mid S_A)$$

By Bayes' rule, the posterior probability of θ_A given S_A, $p(\theta_A \mid S_A)$, can be written as:

$$p(\theta_A \mid S_A) = \frac{p(S_A \mid \theta_A) p(\theta_A)}{p(S_A)}, \quad (4)$$

Assuming a multinomial generation model (Zaragoza et al. 2003) for each summary, parameterized by:

$$\theta_A = (\theta_{A,1}, \theta_{A,2}, ..., \theta_{A,m}) \in [0,1]^m, \quad \sum_{i=1}^{m} \theta_{A,i} = 1.$$

$\theta_{A,i}$ is the parameter of generating word *i* in summary S_A and *m* is the total number of words in the vocabulary. Assuming a bag-of-words unigram model, the system summary likelihood can be expressed as follows:

$$p(S_A \mid \theta_A) = Z_{a_0} \prod_{i=1}^{m} (\theta_{A,i})^{a_i}, \quad (5)$$

where a_i is the number of word *i* occurring in summary S_A, $a_0 = \sum_{i=1}^{m} a_i$, and Z_{a_0} is a constant as:

$$Z_{a_0} = \frac{\Gamma(a_0 + 1)}{\prod_{i=1}^{m} \Gamma(a_i + 1)},$$

where Γ is the gamma function, i.e. $\Gamma(n+1) = n\Gamma(n)$, $\Gamma(0) = 1$, *n* is an integer and $n \geq 0$.

In a MAP estimate, we usually choose the *conjugate* distribution of the generation distribution for a prior. In our case, we assume a Dirichlet prior distribution (the conjugate distribution of the multinomial distribution) as follows:

$$p(\theta_A) = Z'_{\alpha_0} \prod_{i=1}^{m} (\theta_{A,i})^{\alpha_i - 1}, \quad (6)$$

where α_i is hyperparameter related to word *i*, $\alpha_0 = \sum_{i=1}^{m} \alpha_i$, $\alpha_i > 0$, and Z'_{α_0} is:

$$Z'_{\alpha_0} = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{m} \Gamma(\alpha_i)}.$$

By the theory of total probability, the system summary probability can be computed as follows:

$$\begin{aligned} p(S_A) &= \int_\Theta p(S_A \mid \theta_A) p(\theta_A) d\theta_A \\ &= Z_{a_0} Z'_{\alpha_0} \int_\Theta \prod_{i=1}^{m} (\theta_{A,i})^{a_i + \alpha_i - 1} d\theta_A \\ &= Z_{a_0} Z'_{\alpha_0} \frac{\prod_{i=1}^{m} \Gamma(a_i + \alpha_i)}{\Gamma(a_0 + \alpha_0)}. \end{aligned} \quad (7)$$

Substituting (5), (6), and (7) into (4), we have the posterior distribution $p(\theta_A \mid S_A)$ as below:

$$\begin{aligned} & p(\theta_A \mid S_A) \\ &= \frac{p(S_A \mid \theta_A)p(\theta_A)}{p(S_A)} \\ &= \frac{\Gamma(a_0+\alpha_0)}{\prod_{i=1}^{m}\Gamma(a_i+\alpha_i)}\prod_{i=1}^{m}(\theta_{A,i})^{a_i+\alpha_i-1} \\ &= Z'_{a_0+\alpha_0}\prod_{i=1}^{m}(\theta_{A,i})^{a_i+\alpha_i-1}. \end{aligned} \quad (8)$$

We now turn to discuss different ways of estimating $\theta_{A,i}$ and α_i and their implications as described by Zaragoza et al. (2003).

According to Equation (8), the posterior distribution of θ_A given S_A is also a Dirichlet distribution. Its maximum posterior estimation has the following form (Gelman et al. 2003):

$$\theta_{A,i}^{MP} = \frac{a_i+\alpha_i-1}{a_0+\alpha_0-m}, \quad (9)$$

and the posterior distribution (8) can be written as:

$$p(\theta_A \mid S_A) \approx p(\theta_A^{MP} \mid S_A) = Z'_{a_0+\alpha_0}\prod_{i=1}^{m}(\theta_{A,i}^{MP})^{a_i+\alpha_i-1}. \quad (10)$$

If we set $\alpha_i = 1$, then $\theta_{A,i}$ does not depend on α_i, i.e. all possible θ_A's have equal prior. In this case, equation (9) becomes the maximum likelihood estimation as follows:

$$\theta_{A,i}^{ML} = \frac{a_i}{a_0} \quad (11)$$

and the posterior distribution (8) can be written as:

$$p(\theta_A \mid S_A) \approx p(\theta_A^{ML} \mid S_A) = Z'_{a_0+m}\prod_{i=1}^{m}(\theta_{A,i}^{ML})^{a_i}. \quad (12)$$

The problem with using maximum likelihood estimation is when a_i equal to zero. If zero occurrence happens for word *i*, then its maximum likelihood estimation, $\theta_{A,i}^{ML}$, would be zero and the whole posterior distribution would be zero. To tackle this problem, we need to redistribute some probability mass to zero occurrence events or unseen word events. The process of redistribution is called *smoothing* in the language modeling literatures. For an in-depth discussion of this topic, please see Chen and Goodman (1996).

By choosing different value for α_i, we could derive different smoothing methods as discussed in Zaragoza et al. (2003). For example, we could estimate α_i using topic collection frequency by setting $\alpha_i = \mu p(w_i \mid T)+1$, where μ is a scaling factor and $p(w_i \mid T)$ is the probability of word *i* occurring in topic *T*. This is called Bayes-smoothing and has been used in language modeling for information retrieval (Zhai and Lafferty 2004). The Bayes-smoothing can be written as:

$$\theta_{A,i}^{BS} = \frac{a_i+\mu p(w_i \mid T)}{a_0+\mu} \quad (13)$$

Using Equation (13), Equation (8) becomes:

$$p(\theta_A \mid S_A) \approx p(\theta_A^{BS} \mid S_A) = Z'_{a_i+\mu+m}\prod_{i=1}^{m}(\theta_{A,i}^{BS})^{a_i+\mu p(w_i \mid T)}. \quad (14)$$

We now turn to estimating the posterior distribution of θ_R given a reference summary S_R.

2.2 Estimation of Reference Summary Distributions

Given a reference summary S_R, we could estimate posterior distribution θ_R in the same way that we estimate posterior distribution θ_A as follows:

$$\begin{aligned} p(\theta_R \mid S_R) &= \frac{p(S_R \mid \theta_R)p(\theta_R)}{p(S_R)} \\ &= Z'_{a_0+\alpha_0}\prod_{i=1}^{m}(\theta_{R,i})^{a_i+\alpha_i-1}, \end{aligned} \quad (15)$$

where a_i is the number of occurrence of word *i* in reference summary S_R.

Given another reference summary S'_R, i.e., when multiple reference summaries are available, the posterior distribution can be updated using Bayesian inference as follows:

$$\begin{aligned} & p(\theta_R \mid S_R, S'_R) \\ &= \frac{p(\theta_R, S'_R \mid S_R)}{p(S'_R \mid S_R)} \\ &= \frac{p(S'_R \mid \theta_R, S_R)p(\theta_R \mid S_R)}{p(S'_R \mid S_R)} \\ &= \frac{p(S'_R \mid \theta_R)p(\theta_R \mid S_R)}{p(S'_R)} \\ &= \frac{\left(Z_{a'_0}\prod_{i=1}^{m}(\theta_{R,i})^{a'_i}\right)\left(Z'_{a_0+\alpha_0}\prod_{i=1}^{m}(\theta_{R,i})^{a_i+\alpha_i-1}\right)}{Z_{a'_0}Z'_{a_0+\alpha_0}\dfrac{\prod_{i=1}^{m}\Gamma(a'_i+a_i+\alpha_i)}{\Gamma(a'_0+a_0+\alpha_0)}} \\ &= \frac{\Gamma(a'_0+a_0+\alpha_0)}{\prod_{i=1}^{m}\Gamma(a'_i+a_i+\alpha_i)}\prod_{i=1}^{m}(\theta_{T,i})^{a'_i+a_i+\alpha_i-1} \\ &= Z'_{a'_0+a_0+\alpha_0}\prod_{i=1}^{m}(\theta_{R,i})^{a'_i+a_i+\alpha_i-1}, \end{aligned} \quad (16)$$

where $p(\theta_R \mid S_R)$ is the posterior distribution from equation (15), S'_R is independent of S_R, and $p(S'_R)$ is computed using equation (7) but with the posterior distribution $p(\theta_R \mid S_R)$ as prior. In a more general case, given multiple (*L*) reference summaries, $S_{R,1}, \cdots, S_{R,L}$, the posterior distribution of θ_R could

be written as follows by repeat application of Bayesian inference with equation (16):

$$p(\theta_R \mid S_{R,1}, \cdots, S_{R,L}) = Z'_{\alpha_0 + \sum_{j=1}^{L} a_{0,j}} \prod_{i=1}^{m} (\theta_{R,i})^{\alpha_i - 1 + \sum_{j=1}^{L} a_{i,j}}, \quad (17)$$

where $a_{i,j}$ is the number of occurrence of word *i* in reference summary $S_{R,j}$, and

$$a_{0,j} = \sum_{i=0}^{m} a_{i,j}. \quad (18)$$

Equation (18) is the total number of words in reference summary $S_{R,j}$. The total number of words in the topic collection could be computed as follows:

$$\sum_{j=1}^{L} a_{0,j} = \sum_{j=1}^{L} \sum_{i=1}^{m} a_{i,j}. \quad (19)$$

Equation (17) indicates that estimation of posterior distribution given multiple summaries is the same as estimation of posterior distribution given a single summary that contains all the reference summaries. This is reasonable since we assume a bag-of-word unigram likelihood model for generating summaries[4]. It also bodes well with the consensus-oriented manual summarization evaluation approaches proposed by van Halteren and Teufel (2003) and Nenkova and Passonneau (2004). With equations (8) and (17), the summary score of system summary, S_A, can be computed using Jensen-Shannon divergence from equation (3) as follows:

$$Score^{JS}_{summary}(S_A \mid S_R^{1,L}) = -JS_{1/2}(p(\theta_A \mid S_A) \,\square\, p(\theta_R \mid S_R^{1,L})), \quad (20)$$

where $S_R^{1,L}$ is a shorthand for $S_{R,1}, \cdots, S_{R,L}$.

3 Experimental Setup

We used data from DUC 2002 100-word single and multi-document tasks as our testing corpus. DUC 2002 data includes 59 topic sets. Each topic set contains about 10 news article pertaining to some news topic, for example, topic D061 is about "*Hurricane Gilbert*". Two human written summaries per topic are provided as reference summaries. 14 sets of system summaries and 1 simple lead baseline summary are included for the single document summarization task (total 15 runs); while 8 sets of system summaries, 1 lead baseline, and 1 latest news baseline are included for the multi-document summarization task (total 12 runs). All summaries are about 100 words[5]. Manually evaluation results in average coverage[6] scores are also included in the DUC 2002 data set.

The commonly used criterion of success to evaluate an automatic evaluation method is to compute the correlation between the ranking of systems according to human assigned scores and the ranking according to automatic scores (Papineni et al. 2002; Lin & Hovy 2003). We followed the same convention and computed Pearson's product moment correlation coefficient and Spearman's rank correlation coefficient as indicators of success.

Besides evaluating the performance of the automatic evaluation measure based on Jensen-Shannon (*JS*) divergence as defined in equation (2), we also compared it with measures based on *KL*-divergence and simple log likelihood. The effect of smoothing and the difference of using single and multiple reference summaries were also investigated. To examine the effect of using longer n-grams ($n > 1$), we also used bag-of-bigram and bag-of-trigram models by simply replace unigrams in the model proposed in Section 2 with bigrams and trigrams and treat them as unigrams. Lemur toolkit version 4.0[7] was used to estimate models with modification to speedup computation of bigram and trigram models. We also ran standard ROUGE v1.5.5 with ROUGE1 to 4 as baselines.

All experiments were run with common words excluded and Porter stemmer applied. We summarize these experiments in the following sections.

3.1 Jensen-Shannon Divergence (JSD)

We use equation (22) to compute summary score and apply maximum likelihood estimation (θ^{ML}) of the parameters according to equation (11). Using a unigram model and single reference summary, we rewrite equation (22) as follows:

[4] Please refer to equation (5).

[5] There were also 10-, 50-, and 200-word summary tasks in DUC 2002 multi-document summarization evaluation. However, we only used the data of 100-word summarization subtask for this experiment.

[6] Coverage is a weighted recall metric measuring the overlap of content between a system summary and a reference summary. For example, if 4 elementary discourse units (EDU, Marcu 1998) in a system summary partially match EDUs in a reference summary of total 8 EDUs, then its coverage score is R*4/8. R is the ratio of the partial match. R is 20%, 40%, 60%, 80%, or 100% in DUC 2002. A single human assigned the ratio using only one reference. The other reference was not used in the manual evaluation.

[7] The Lemur project: http://www.lemurproject.org.

$$Score^{JSD}_{summary}(S_A \mid S_R^{1,1}) =$$

$$-\frac{1}{2}\sum_{\theta_i^{MP}}\left(\begin{array}{l} p(\theta_i^{MP} \mid S_A)\log\left(\dfrac{p(\theta_i^{MP} \mid S_A)}{\frac{1}{2}p(\theta_i^{MP} \mid S_A)+\frac{1}{2}p(\theta_i^{MP} \mid S_R^{1,1})}\right)+ \\ p(\theta_i^{MP} \mid S_R^{1,1})\log\left(\dfrac{p(\theta_i^{MP} \mid S_R^{1,1})}{\frac{1}{2}p(\theta_i^{MP} \mid S_A)+\frac{1}{2}p(\theta_i^{MP} \mid S_R^{1,1})}\right)\end{array}\right)$$

where $p(\theta_i^{MP} \mid S_A)$ and $p(\theta_i^{MP} \mid S_R^{1,1})$ are estimated as follows:

$$\begin{aligned} p(\theta_i^{MP} \mid S_A) &= \frac{a_{A,i}}{a_{A,0}} \\ &= \frac{C(w_i,S_A)}{\sum_{w_i} C(w_i,S_A)}, \\ p(\theta_i^{MP} \mid S_R^{1,1}) &= \frac{a_{R,i}}{a_{R,0}} \\ &= \frac{C(w_i,S_R^{1,1})}{\sum_{w_i} C(w_i,S_R^{1,1})}. \end{aligned}$$

$C(w_i,S_A)$ and $C(w_i,S_R^{1,1})$ are the counts of word w_i in system summary S_A and reference summary $S_R^{1,1}$ respectively. When multiple reference summaries are used, $p(\theta_i^{MP} \mid S_R^{1,L})$ is estimated as follows:

$$\begin{aligned} p(\theta_i^{MP} \mid S_R^{1,L}) &= \frac{a_{R,i}^{1,L}}{a_{R,0}^{1,L}} \\ &= \frac{\sum_j C(w_i,S_{R,i}^{j,j})}{\sum_{w_i}\sum_j C(w_i,S_{R,i}^{j,j})}. \end{aligned}$$

3.2 Jensen-Shannon Divergence with Smoothing (JSDS)

To examine the effect of smoothing when we compute summary score using equation (22), we apply Bayes-smoothing as shown in equation (15). Using a unigram model and single reference summary, we rewrite equation (22) as follows:

$$Score^{JSDS}_{summary}(S_A \mid S_R^{1,1}) =$$

$$-\frac{1}{2}\sum_{\theta_i^{BS}}\left(\begin{array}{l} p(\theta_i^{BS} \mid S_A)\log\left(\dfrac{p(\theta_i^{BS} \mid S_A)}{\frac{1}{2}p(\theta_i^{BS} \mid S_A)+\frac{1}{2}p(\theta_i^{BS} \mid S_R^{1,1})}\right)+ \\ p(\theta_i^{BS} \mid S_R^{1,1})\log\left(\dfrac{p(\theta_i^{BS} \mid S_R^{1,1})}{\frac{1}{2}p(\theta_i^{BS} \mid S_A)+\frac{1}{2}p(\theta_i^{BS} \mid S_R^{1,1})}\right)\end{array}\right)$$

where $p(\theta_i^{BS} \mid S_A)$ and $p(\theta_i^{BS} \mid S_R^{1,1})$ are estimated as follows:

$$\begin{aligned} p(\theta_i^{BS} \mid S_A) &= \frac{a_{A,i}+\mu p(w_i \mid C)}{a_{A,0}+\mu} \\ &= \frac{C(w_i,S_A)+\mu p(w_i \mid C)}{\left(\sum_{w_i} C(w_i,S_A)\right)+\mu}, \\ p(\theta_i^{BS} \mid S_R^{1,1}) &= \frac{a_{R,i}+\mu p(w_i \mid C)}{a_{R,0}+\mu} \\ &= \frac{C(w_i,S_R^{1,1})+\mu p(w_i \mid C)}{\left(\sum_{w_i} C(w_i,S_R^{1,1})\right)+\mu}. \end{aligned}$$

$C(w_i,S_A)$ and $C(w_i,S_R^{1,1})$ are the counts of word w_i in system summary S_A and reference summary $S_R^{1,1}$ respectively. The Bayes-smoothing probability or Bayesian prior $p(w_i \mid C)$ is estimated from a general English corpus instead of the topic collection as we described in section 2.1. In our experiments, we used TREC AP88-90 collection that contained more than 200,000 news articles. When multiple reference summaries are used, $p(\theta_i^{BS} \mid S_R^{1,L})$ is estimated as follows:

$$\begin{aligned} p(\theta_i^{BS} \mid S_R^{1,L}) &= \frac{a_{R,i}^{1,L}+\mu p(w_i \mid C)}{a_{R,0}^{1,L}+\mu} \\ &= \frac{\left(\sum_j C(w_i,S_{R,i}^{j,j})\right)+\mu p(w_i \mid C)}{\left(\sum_{w_i}\sum_j C(w_i,S_{R,i}^{j,j})\right)+\mu}. \end{aligned}$$

The value of μ could be determined empirically. In this experiment we set μ to 2,000 following Zhai and Lafferty (2004).

3.3 Kullback-Leibler Divergence with Smoothing (KLDS)

To compare the performance of *JSD* and *JSDS* scoring methods with other alternative distance measure, we also compute summary scores using *KL* divergence with Bayes-smoothing as follows:

$$Score^{KL}_{summary}(S_A \mid S_R^{1,L}) = -\sum_{\theta_i^{BS}} p(\theta_i^{BS} \mid S_A)\log\left(\frac{p(\theta_i^{BS} \mid S_A)}{p(\theta_i^{BS} \mid S_R^{1,L})}\right)$$

The Bayes-smoothing factor μ is also set to 2,000 and θ_i^{BS} is estimated by the same way that we compute *JSDS*.

Unigram		JSD		JSDS		KLDS		LLS	
		P	S	P	S	P	S	P	S
SD	SR	**0.97**	**0.91**	0.61	0.25	0.59	0.23	-0.54	0.16
	MR	**0.97**	**0.91**	0.62	0.65	0.61	0.25	-0.60	0.11
MD	SR	**0.80**	**0.83**	0.44	0.64	0.34	0.54	0.21	0.36
	MR	**0.88**	**0.89**	0.76	0.81	0.61	0.71	0.47	0.60

Bigram		JSD		JSDS		KLDS		LLS	
		P	S	P	S	P	S	P	S
SD	SR	**0.92**	**0.90**	0.64	0.60	0.50	0.20	-0.81	0.05
	MR	**0.94**	**0.90**	0.64	0.62	0.53	0.26	-0.80	0.06
MD	SR	**0.91**	**0.88**	-0.17	-0.19	0.01	0.14	0.82	0.87
	MR	**0.96**	**0.94**	0.17	0.36	0.12	0.22	0.85	0.89

Trigram		JSD		JSDS		KLDS		LLS	
		P	S	P	S	P	S	P	S
SD	SR	**0.92**	**0.90**	0.68	0.44	0.53	0.11	-0.72	0.03
	MR	**0.94**	**0.90**	0.68	0.58	0.55	0.20	-0.71	0.01
MD	SR	**0.87**	**0.82**	-0.39	-0.33	-0.11	-0.10	0.50	0.54
	MR	**0.93**	**0.89**	-0.30	-0.26	-0.11	-0.10	0.54	0.54

Table 1. DUC 2002 single (SD) and multi-document summarization (MD) tasks' Pearson's (P) and Spearman's (S) correlations of automatic measures (*JSD*, *JSDS*, *KLDS*, and *LLS*) using single (SR) or multiple (MR) reference summaries. (Unigram: bag-of-unigram model, Bigram: bag-of-bigram model, and Trigram: bag-of-trigram model)

3.4 Log Likelihood with Smoothing (LLS)

As a baseline measure, we also compute the log likelihood score of an automatic summary given a reference summary or a set of reference summaries as follows:

$$Score^{LL}_{summary}(S_A \mid S_R^{1,L}) = \sum_{i=1}^{|S_A|} \log p(\theta_i^{BS} \mid S_R^{1,L}),$$

where $|S_A|$ is the length of S_A and $p(\theta_i^{BS} \mid S_R^{1,L})$ is estimated as before.

4 Results

Table 1 shows the results of all runs. According to Table 1, automatic evaluation measure based on Jensen-Shannon Divergence without Bayes-smoothing (*JSD*) performed the best among all measures. Among them, *JSD* over the bag-of-unigram model achieved the best results in the single document summarization task (P-SD-MR: 0.97, S-SD-MR: 0.91); while the bag-of-bigram model achieved the best results in the multiple document summarization task (P-MD-MR: 0.96, S-MD-MR: 0.94). Although the bag-of-bigram model did not perform as well as the bag-of-unigram model in the single document summarization task, its Pearson (SD-MR: 0.94) and Spearman (SD-MR: 0.90) correlation values were still over 90% regardless of single or multiple references were used.

ROUGE		ROUGE-1		ROUGE-2		ROUGE-3		ROUGE-4	
		P	S	P	S	P	S	P	S
MR	SD	0.99	0.84	1.00	0.96	1.00	0.98	**1.00**	**0.99**
	MD	0.70	0.59	0.89	0.84	**0.92**	**0.85**	0.90	0.78

Table 2. DUC 2002 single (SD) and multi-document (MD) summarization tasks' Pearson's (P) and Spearman's (S) correlations of automatic measures (ROUGE1-4) using multiple (MR) reference summaries.

We also observed that using multiple references outperformed using only single reference. This is reasonable since we expect to estimate models better when more reference summaries are available.

Smoothed measures did not perform well. This is not a surprise due to the nature of summarization evaluation. Intuitively, only information presented in system and reference summaries should be considered for evaluation.

The *JSD*-based measure was also compared favorably to ROUGE in the multiple document summarization task as shown in Table 2. In particular, the *JSD*-based measure over bag-of-bigram model using multiple references achieved much better results in both Pearson's and Spearman's correlations than all versions of ROUGE. For single document summarization task, the *JSD*-based measure still achieved high correlations (90%+) though it was not as high as ROUGE2, 3, and 4.

5 Related Work

The approach described in this paper is most similar to the Bayesian extension in information retrieval (IR) work by Zaragoza et al. (2003). In their work, query likelihood model was presented as Bayesian inference. Other earlier language modeling (Rosenfeld 2002) work in information retrieval, especially the idea of modeling a document using bag-of-word unigram model, also inspire this work (Berger and Lafferty 1999, Lafferty and Zhai 2001

Statistical language models such as document language model (Ponte and Croft 1998, Zhai and Lafferty 2004), relevance-based language models (Lavrenko and Croft 2001), and dependency-based language models (Gao et al. 2004) have been applied successfully in information retrieval. It has also been applied to topic detection and tracking (Lavrenko et al. 2002, Larkey et al. 2004). Ex-

tended models also have been developed to deal with vocabulary mismatch and query expansion problems (Berger and Lafferty 1999, Hofmann 1999, Lafferty and Zhai 2001). However, it has not been applied in automatic evaluation of summarization. Hori et al. (2004) also considered using "posterior probability" derived from consensus among human summaries as weighting factor to improve evaluations of speech summarization. But their notion of "posterior probability" was not true probability and was not presented as an integral part of the Bayesian inference framework as we have described in this paper.

6 Conclusions and Future Work

The research proposed in this paper aims at providing a pilot study of applying information-theoretic measures in automatic evaluation of summaries. With the initial success of this study, we would like to: (1) verify the results with other set of data, for example, DUC 2003 data, (2) tune the Bayesian smoothing parameter μ to further examine the effect of smoothing, (3) develop better content generation model and (4) add synonym and paraphrase matching capability in the future. To address (3), for example, we would like to explore mutual information-based dependency language modeling as proposed by Gao et al. (2004).

For (4), manual evaluation methods recently proposed separately by van Halteren and Teufel (2003), the factoid method, and Nenkova and Passouneau (2004), the pyramid method, tried to take advantage of the availability of multiple references. Both methods assume that the more important a piece of information is, the more reference summaries it appears in. These manual evaluation methods can identify semantic equivalents. For example, a summary content unit (SCU) "The diamonds were replaced by fake replicas[8]" created as defined in Nenkova and Passouneau (2004) from the following four contributing clauses (1a – d):

1. Authorities, responding to a tip, [*switched the diamonds with fakes*]$_{1a}$ and were waiting inside the building dressed as cleaners when the thieves burst in with a bulldozer and sledge-hammers.
2. However, authorities were tipped off and [*switched the diamonds with fakes*] $_{1b}$.
3. They disguised themselves as cleaners at the Millennium Dome, [*switched the diamonds with worthless glass*] $_{1c}$, and waited for the robbers, who planned to get away in a speedboat down the Thames River.
4. [*The diamonds had been swapped with glass replicas*] $_{1d}$.

Contributors (1a – d) from 4 reference summaries to the SCU are underlined. The manual pyramid method can identify these semantic equivalents. It is obvious that automatic evaluation methods relying on strict n-gram or lexical matching would only find two out of four possible matches, i.e. "*switched the diamonds with fakes*" from (1a) and (1b) while leave "*switched the diamonds with worthless glass*" (1c) and "*The diamonds had been swapped with glass replicas*" (1d) unmatched. Allowing near synonyms such as *fakes*, *worthless glass*, and *glass replicas* to match might help, but how to acquire these equivalents and how to assign appropriate weights to reflect their subtle differences remain open questions. To find semantic equivalents automatically, we would like to try query expansion techniques (Hofmann 1999, Lafferty and Zhai 2001, Bai et al. 2005, Cao et al. 2005) commonly used in IR. Proper query expansion boosts IR system performance. We suspect that these techniques would help a little but we probably would need to develop much better paraphrase expansion and matching techniques to see significant boost in overall performance.

[8] Example is taken from Multilingual Summarization Evaluation 2005 (MSE2005), topic number 33003.

7 Acknowledgement

Part of this work was conducted while the first author visited Microsoft Research Asia (MSRA) in the summer of 2005. He would like to thank Ming Zhou, Hsiao-Wuen Hon, and other staffs at MSRA for providing an excellent research environment and exciting intellectual exchanges during his visit.

Reference

Bai, Jing, Dawei Song, Peter Bruza, Jian-Yun Nie, and Guihong Cao. 2005. Query Expansion Using Term Relationships in Language Models for Information Retrieval. *Proceedings of International Conference on Information and Knowledge Management (CIKM) 2005*. October 31 – November 5, 2005, Bremen, Germany.

Berger, Adam and John Lafftery. 1999. *Information Retrieval as Statistical Translation*. Proceedings of ACM-SIGIR 1999. August 15-19, 1999, Berkeley, CA, USA.

Cao, Guihong, Jian-Yun Nie, and Jing Bai. 2005. Integrating Word Relationships into Language Models. *Proceedings of SIGIR 2005*. August 15-19, 2005, Salvador, Brazil.

Chen, Stanley and Joshua Goodman. 1996. An Empirical Study of Smoothing Techniques for Language Modeling. *Proceedings of 34th Annual Meeting on Association for Computational Linguistics*, page 310-318, June 23-28, Santa Cruz, California, USA.

Dagan, Ido, Lillian Lee, and Fernado C. N. Pereira. 1999. Similarity-Based Models of Word Cooccurrence Probability. *Machine Learning*. Vol 34, page 43-69, 1999.

Gao, Jianfeng, Jian-Yun Nie, Guangyuan Wu, and Guihong Cao. 2004. Dependence Language Model for Information Retrieval. *Proceedings of SIGIR 2004*. July 25-29, 2004, Sheffield, UK.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2003. *Bayesian Data Analysis*. 2nd Edition. Chapman & Hall/CRC.

Hofmann, Thomas. 1999. Probabilistic Latent Semantic Indexing. *Proceedings of ACM-SIGIR 1999*. August 15-19, 1999, Berkeley, CA, USA.

Hori, Chiori, Tsutomu Hirao, and Hideki Isozaki. 2004. Evaluation Measures Considering Sentence Concatenation for Automatic Summarization by Sentence or Word Extraction. *Proceedings of Workshop on Text Summarization Branches Out*. July 25, 2004, Barcelona, Spain.

Kraaij, Wessel, Martijn Spitters, and Martine van der Heijden. 2001. Combining a Mixture Language Model and Naïve Bayes for Multi-Document Summarisation. *Proceedings of DUC 2001*.

Lafferty, John and Chiangxiang Zhai. 2001. Document Language Models, Query Models, and Risk Minimization for Information Retrieval. *Proceedings of ACM-SIGIR 2001*. September 9-13, 2001, New Orleans, LA, USA.

Larkey, Leah S., Fangfang Feng, Margaret Connell, and Victor Larvrenko. 2004. Language-specific Models in Multilingual Topic Tracking. *Proceedings of ACM-SIGIR 2004*. July 25-29, 2004. Sheffield, UK.

Lavrenko, Victor, James Allan, Edward DeGuzman, Daniel LaFlamme, Veera Pollard, and Stephen Thomas. 2002. Relevance Models for Topic Detection and Tracking. *Proceedings of HLT 2002*. March 24-27, 2002, San Diego, CA, USA.

Lavrenko, Victor and W. Bruce Croft. 2001. Relevance-Based Language Models. *Proceedings of ACM-SIGIR 2001*. September 9-13, 2001, New Orleans, LA, USA.

Lin, Chin-Yew and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. *Proceedings of HLT-NAACL-2003*. May 27-June 1, 2003, Edmonton, Canada.

Lin, Chin-Yew. 2004. ROUGE: a Package for Automatic Evaluation of Summaries. *Proceedings of Workshop on Text Summarization 2004*. July 21-26, 2004, Barcelona, Spain.

Lin, Jianhua. 1991. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37(1), page 145-151, 1991.

Nenkova, Ani and Rebecca Passonneau. 2004. Evaluating Content Selection in Summarization: the Pyramid Method. *Proceedings of NAACL-HLT 2004*. May 2-7, 2004, Boston, MA, USA.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, Philadelphia, USA, July 2002, page 311-318.

Ponte, Jay M. and W. Bruce Croft. 1998. A language modeling approach to information retrieval. *Proceedings of ACM-SIGIR 1998*, pages 275-281. August 24-28, 1998, Melbourne, Australia.

Rosenfeld, Ronald. 2002. Two Decades of Statistical Language Modeling: Where do We Go from Here? *Proceedings of IEEE*.

Van Halteren, Hans and Simone Teufel. 2003. Examining the Consensus between Human Summaries: Initial Experiments with Factoid Analysis. *Proceedings of Workshop on Text Summarization 2003*. May 27-June 1, 2003, Edmonton, Canada.

Zaragoza, Hugo, Djoerd Hiemstra, and Michael Tipping. 2003. Bayesian Extension to the Language Model for Ad Hoc Information Retrieval. *Proceedings of ACM-SIGIR 2003*. July 28-August 1, 2003, Toronto, Canada.

Zhai, Chengxiang and John Lafferty. 2004. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems*, Vol 22, No. 2, April 2004, pages 179-214.