SPOKEN DOCUMENT SUMMARIZATION USING RELEVANT INFORMATION

Yi-Ting Chen[1], Shih-Hsiang Lin[2], Hsin-Min Wang[1] and Berlin Chen[2]

[1]Institute of Information Science, Academia Sinica, Taiwan
[2]National Taiwan Normal University, Taiwan
{ytchen, whm}@iis.sinica.edu.tw, {69308027, berlin}@ntnu.edu.tw

ABSTRACT

Extractive summarization usually automatically selects indicative sentences from a document according to a certain target summarization ratio, and then sequences them to form a summary. In this paper, we investigate the use of information from relevant documents retrieved from a contemporary text collection for each sentence of a spoken document to be summarized in a probabilistic generative framework for extractive spoken document summarization. In the proposed methods, the probability of a document being generated by a sentence is modeled by a hidden Markov model (HMM), while the retrieved relevant text documents are used to estimate the HMM's parameters and the sentence's prior probability. The results of experiments on Chinese broadcast news compiled in Taiwan show that the new methods outperform the previous HMM approach.

Index Terms— extractive summarization, hidden Markov model, probabilistic generative model, relevant document, relevance model, spoken document summarization

1. INTRODUCTION

The vast amount of multimedia content that enriches our daily lives continues to grow at a phenomenal rate. Speech is one of the most important sources of information about multimedia content, since it usually represents the concepts and topics of the content. As a result, multimedia access based on associated spoken documents has received a great deal of attention in recent years. However, unlike text documents, which are usually structured with titles and paragraphs, and are therefore easy to retrieve and browse, spoken documents are only represented by audio signals and are difficult to browse. Though spoken documents can be automatically transcribed into a word format, incorrect recognition results and redundant acoustic effects make accessing them difficult. Spoken document summarization, which tries to distil important information and remove redundant and incorrect content, can help users review spoken documents efficiently and understand associated topics quickly [1].

This paper investigates extractive spoken document summarization, which automatically selects indicative sentences from a document according to a certain target summarization ratio, and then sequences them to form a summary. Existing extractive summarization approaches generally fall into three categories: 1) approaches based on the sentence structure or location information, 2) approaches based on statistical measures, and 3) approaches based on the generative probability. In [2, 3], the authors suggest that important sentences can be selected from the significant parts of a document, e.g., the introduction and conclusion. However, such approaches can only be applied to documents in some specific domains or documents that have some specific structures. Statistical approaches that select salient sentences based on the statistical features, such as the term (word) frequencies, language model scores, acoustic confidence measures, and prosodic information, of the sentences or the words in the sentences have attracted much attention in recent years. Representative methods in this category include the vector space model (VSM) [4], the latent semantic analysis (LSA) method [5], the maximum marginal relevance (MMR) method [6], and the sentence significant score method [3, 7]. In addition, several classification-based methods using statistical features have been developed, for example, the Bayesian network classifier [8], the support vector machine (SVM) [9, 10], and the logistic regression model [9]. In these methods, sentence selection is formulated as a binary classification problem; however, a training set comprising documents and their corresponding handcrafted summaries (or labeled data) is needed to train the classifiers. Recently, several approaches based on the probabilistic generative model have also been proposed. The hidden Markov model (HMM) [11], the sentence topical mixture model (STMM) [12], and the word topical mixture model (WTMM) [13, 14] have all demonstrated competitive results in the Chinese spoken document summarization task.

In this paper, we attempt to improve extractive spoken document summarization by using information from relevant documents for each sentence of a spoken document to be summarized in a HMM-based probabilistic generative framework. Because there is no prior knowledge about the relevant set for each sentence, a local feedback procedure

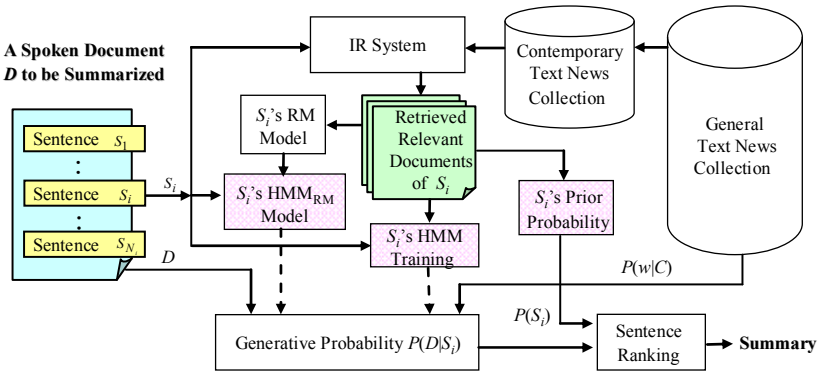

Figure 1: A schematic depiction of a probabilistic generative framework for extractive spoken document summarization using information from documents retrieved from a text collection.

[15] is employed by taking the sentence as a query and sending it to the information retrieval (IR) system to obtain a ranked list of documents from a large text collection. It is assumed that the top K documents returned by the IR system are relevant to the sentence, and are thus treated as the relevant set of the sentence. Local feedback was originally applied in IR [15], where the top ranked documents retrieved by the original query were assumed relevant to the query, and used to re-weight and expand the query [15], or construct a relevance model for the query [16, 17]. In the proposed methods, the probability of a document being generated by a sentence is modeled by a HMM, while the retrieved relevant text documents are used to estimate the HMM's parameters and the sentence's prior probability. The results of experiments on Chinese broadcast news compiled in Taiwan show that the new methods achieve noticeable performance gains over the previous HMM approach [11].

The remainder of this paper is organized as follows. Section 2 reviews the HMM-based probabilistic generative framework for extractive spoken document summarization. Section 3 explains how we use information from relevant documents retrieved for a sentence of a spoken document in the HMM-based probabilistic generative framework. The experiment setup and the results are discussed in Sections 4 and 5, respectively. Finally, in Section 6, we present our conclusions.

2. THE PROBABILISTIC GENERATIVE FRAMEWORK FOR EXTRACTIVE SPOKEN DOCUMENT SUMMARIZATION

In the probabilistic generative framework for extractive spoken document summarization, the importance of a sentence S_i in a document D can be modeled by $P(S_i|D)$; i.e., the posterior probability of the sentence S_i given the document D. According to Bayes' rule, $P(S_i|D)$ can be expressed as:

$$P(S_i|D) = \frac{P(D|S_i)P(S_i)}{P(D)}, \quad (1)$$

where $P(D|S_i)$ is the generative probability of the document D given the sentence S_i; $P(S_i)$ is the prior probability of S_i being important; and $P(D)$ is the prior probability of D. In Eq. (1), $P(D)$ can be eliminated because it is identical for all sentences and will not affect their ranking. The generative probability $P(D|S_i)$ can be considered as a relevance measure between the document D and the sentence S_i, while the sentence's prior probability $P(S_i)$ is, to some extent, a measure of the importance of the sentence itself. Therefore, all the sentences of the spoken document to be summarized are ranked according to the product of the generative probability $P(D|S_i)$ and the sentence's prior probability $P(S_i)$. Then, the sentences with the highest probabilities are selected and sequenced to form a summary.

2.1. Hidden Markov Model (HMM)

In our previous work [11], HMM was applied to extractive spoken document summarization, where each sentence S_i of a document D to be summarized was treated as a probabilistic generative model (or HMM) consisting of n-gram distributions for predicting the document D, and the terms (or words) in the document D were taken as an input observation sequence. It was assumed that the sentence's prior probability $P(S_i)$ was uniformly distributed. When only the unigrams were considered, the generative probability of the document D given the sentence S_i was expressed as

$$P_{\text{HMM}}(D|S_i) = \prod_{w \in D} \left[\lambda \cdot P(w|S_i) + (1-\lambda) \cdot P(w|C)\right]^{n(w,D)}, \quad (2)$$

where λ was a weighting parameter and $n(w,D)$ was the occurrence count of the term w in D. The sentence model $P(w|S_i)$ and the collection model $P(w|C)$ were estimated, respectively, from the sentence itself and a large external text collection using the maximum likelihood estimation

(MLE). Although a sentence-dependent weighting parameter λ in Eq. (2) can be trained by taking the document D as the training observation sequence and using the EM training formula

$$\hat{\lambda} = \frac{\sum_{w \in D} n(w, D) \cdot \frac{\lambda \cdot P(w \mid S_i)}{\lambda \cdot P(w \mid S_i) + (1-\lambda) \cdot P(w \mid C)}}{\sum_{w \in D} c(w, D)}, \quad (3)$$

the sentence-dependent weighting parameter λ obtained in this way is not reliable because the training data is sparse and there could be errors generated by automatic speech recognition. Therefore, we used a fixed value of 0.05 for λ in our previous work.

Once the HMM for each sentence in the document D was estimated, it was used to predict the occurrence probability of the terms in D. The sentences with the highest probabilities were then selected and sequenced to form the final summary according to different summarization ratios.

3. THE IMPROVED PROBABILISTIC GENERATIVE FRAMEWORK USING RELEVANT INFORMATION

In the HMM approach, as shown in Eq. (2), the sentence model $P(w|S_i)$ is smoothed by the collection model $P(w|C)$. However, the sentence model $P(w|S_i)$ might not be accurately estimated by MLE, since the sentence consists of only a few terms, and the portions of the terms in the sentence are not the same as the probabilities of those terms in the true model. The task becomes even more difficult when there are recognition errors in the spoken sentence. Motivated by the concept of the relevance model used in information retrieval [16, 17], we believe that relevant documents associated with a sentence retrieved from a contemporary text collection could be used to yield a more accurate estimation of the sentence model $P(w|S_i)$ and the weighting parameter λ in Eq. (2), or to estimate the sentence's prior probability $P(S_i)$ in Eq. (1). Figure 1 shows a diagram of a probabilistic generative framework for extractive spoken document summarization using information from documents retrieved from a contemporary text collection.

3.1. Estimation of the Sentence Model in HMM

The first way to use the relevant documents associated with the spoken sentence retrieved from the contemporary text collection is to form a more accurate sentence model $\hat{P}(w \mid S_i)$ by combining the original sentence model $P(w|S_i)$ in Eq. (2) with the relevance model $P(w|R_{S_i})$ estimated from the relevant documents. Each sentence S_i of the document D to be summarized has its own associated relevant set $\{R_{S_i}\}$, which can be approximated by the set of documents retrieved from a large text collection by taking the sentence S_i as a query. Therefore, the relevance model $P(w|R_{S_i})$ of S_i can be constructed by the following equation:

$$P(w \mid R_{S_i}) \approx \sum_{D_r \in \{R_{S_i}\}} P(D_r \mid S_i) P(w \mid D_r), \quad (4)$$

where $P(D_r|S_i)$ can be approximated by the following equation using Bayes' rule:

$$P(D_r \mid S_i) \approx \frac{P(D_r) P(S_i \mid D_r)}{\sum_{D_l \in \{R_{S_i}\}} P(D_l) P(S_i \mid D_l)}. \quad (5)$$

In Eq. (5), $P(D_l)$ is assumed to be uniformly distributed, while $P(S_i|D_l)$ is also modeled by a HMM in the HMM-based information retrieval system [18]. Consequently, the relevance model $P(w|R_{S_i})$ is combined linearly with the original sentence model $P(w|S_i)$ to form a more accurate sentence model:

$$\hat{P}(w \mid S_i) = \alpha \cdot P(w \mid S_i) + (1-\alpha) \cdot P(w \mid R_{S_i}), \quad (6)$$

where α is a weighting parameter. Then, Eq. (2) can be re-written as:

$$P_{\mathrm{HMM_{RM}}}(D \mid S_i) = \prod_{w \in D} \left[\lambda \cdot \hat{P}(w \mid S_i) + (1-\lambda) P(w \mid C) \right]^{n(w,D)}. \quad (7)$$

In this paper, we denote this model as $\mathrm{HMM_{RM}}$.

3.2. Estimation of the Weighting Parameter in HMM

As mentioned in Section 2.1, in our previous work, the weighting parameter λ was fixed because it could not be estimated reliably by using only the spoken document to be summarized. Since the relevant documents retrieved for the spoken sentence from the contemporary text collection are statistically relevant to the spoken sentence, they might be used instead of the spoken document as the training data to estimate λ. Given a set of relevant documents $\{R_{S_i}\}$ for each sentence S_i, the sentence-dependent weighting parameter λ can be estimated by

$$\hat{\lambda} = \frac{\sum_{D_r \in \{R_{S_i}\}} \sum_{w \in D_r} n(w, D_r) \cdot \frac{\lambda \cdot P(w \mid S_i)}{\lambda \cdot P(w \mid S_i) + (1-\lambda) \cdot P(w \mid C)}}{\sum_{D_r \in \{R_{S_i}\}} \sum_{w \in D_r} n(w, D_r)}. \quad (8)$$

In Eq. (3), a single spoken document that could contain incorrect terms is used, whereas Eq. (8) considers multiple text documents; hence, the estimation of the sentence-dependent weighting parameter λ in Eq. (8) is more reliable. We denote this model as $\mathrm{HMM_T}$.

3.3. Estimation of the Sentence's Prior Probability

Because the way to estimate the prior probability of the sentence is still an open issue, researchers usually assume that the sentence's prior probability is uniformly distributed. However, the sentences in a spoken document to be summarized should not be considered equal in importance. In our previous work [14], we attempted to model a sentence's prior probability based on a set of features, such as language model scores, acoustic confidence measures, and prosodic information, extracted from the spoken sentences. However, the sentence's prior probability might not be accurately estimated by the above features, since there are recognition errors, incorrect boundaries, and redundant information.

In this study, we observed that the retrieved text documents for a spoken sentence with more key words or correctly recognized words usually have the same or similar topics. On the other hand, the retrieved documents for a spoken sentence with more common words or recognition errors usually have diverse topics. Therefore, the similarity among the retrieved text documents might be a useful indicator of the importance of the spoken sentence, and the prior probability of sentence S_i can be expressed as

$$P(S_i) = \frac{avgSim(S_i)}{\sum_{S \in D} avgSim(S)}, \tag{9}$$

where $avgSim(S_i)$ is the average similarity of documents in the relevant set $\{R_{S_i}\}$ for the sentence S_i computed by

$$avgSim(S_i) = \frac{\sum_{D_r \in \{R_{S_i}\}} \sum_{\substack{D_x \in \{R_{S_i}\} \\ D_x \neq D_r}} \frac{\overrightarrow{D_r} \cdot \overrightarrow{D_x}}{\left|\overrightarrow{D_r}\right| \cdot \left|\overrightarrow{D_x}\right|}}{\left|R_{S_i}\right| \cdot \left(\left|R_{S_i}\right| - 1\right)}, \tag{10}$$

where $\overrightarrow{D_x}$ is the tf-idf (term frequency – inverse document frequency) vector representation of the document D_x, and $\left|R_{S_i}\right|$ is the number of documents in the relevant set $\{R_{S_i}\}$. We denote this model as HMM-P.

4. EXPERIMENT SETUP

4.1. Speech and Text Corpora

The speech data set was comprised of approximately 176 hours of radio and TV broadcast news documents collected from several radio and TV stations in Taipei between 1998 and 2004. From them, a subset of 200 documents (1.6 hours) collected in August 2001 was reserved for the summarization experiments [4]. The Chinese character error rate (CER) was 14.17%. The 200 broadcast news documents were divided into two parts, each containing 100 spoken documents. The first part was taken as the development set, which formed the basis for tuning the parameters or settings. The second part was taken as the evaluation set; i.e., all the summarization experiments conducted on it followed the same training (or parameter) settings, which were optimized based on the development set.

A large number of text news documents collected from the Central News Agency (CNA) between 1991 and 2002 (the Chinese Gigaword Corpus released by LDC) was also used [19]. The text news documents collected in 2000 and 2001 were used to train *n*-gram language models for speech recognition and the collection model $P(w|C)$ in Eq. (2); and a subset of about 14,000 text news documents collected in August 2001 was used as the contemporary text collection to construct relevant information.

4.2. Evaluation Metric

Three subjects were asked to summarize the 200 broadcast news documents to be used as references for evaluation. The ROUGE measure [20] was used to evaluate the performance levels of the proposed models. The measure evaluates the quality of the summarization by counting the number of overlapping units, such as *n*-grams and word sequences, between the automatic summary and a set of reference (or manual) summaries. ROUGE-N is an *n*-gram recall measure defined as follows:

$$ROUGE-N = \frac{\sum_{M \in \mathbf{M}_R} \sum_{gram_n \in M} Count_{match}(gram_n)}{\sum_{M \in \mathbf{M}_R} \sum_{gram_n \in M} Count(gram_n)}, \tag{11}$$

where N denotes the length of the *n*-gram; M is an individual reference (or manual) summary; $\mathbf{M}_R$ is a set of reference summaries; $Count_{match}(gram_n)$ is the maximum number of *n*-grams co-occurring in the automatic summary and the reference summary; and $Count(gram_n)$ is the number of *n*-grams in the reference summary. In this paper, we adopt the ROUGE-2 measure, which uses word bigrams as the matching units.

5. EXPERIMENT RESULTS

5.1. Estimation of HMM's Parameters Using Relevant Information

Tables 1 and 2 show the results of experiments on the development set and the evaluation set, respectively. In the experiments, we used the HMM approach [11] as the baseline system. The comparison of HMM with other summarization models, such as VSM [4], MMR [6], and LSA [5], was reported in [14].

Table 1: The results achieved by different summarization methods under different summarization ratios. (Development set)

	HMM	HMM_{RM}	HMM_T	HMM-P	HMM_{RM}-P	HMM_T-P
10%	0.3084	0.3369	0.3528	0.3696	0.3450	0.3588
20%	0.3467	0.3757	0.3851	0.3954	0.3911	0.3966
30%	0.3734	0.3725	0.3842	0.3796	0.3746	0.3780
50%	0.4768	0.4779	0.4952	0.4814	0.4830	0.4922

Table 2: The results achieved by different summarization methods under different summarization ratios. (Evaluation set)

	HMM	HMM_{RM}	HMM_T	HMM-P	HMM_{RM}-P	HMM_T-P
10%	0.2932	0.3182	0.3316	0.3701	0.3650	0.3654
20%	0.3191	0.3264	0.3412	0.3651	0.3554	0.3582
30%	0.3705	0.3671	0.3739	0.3671	0.3724	0.3864
50%	0.4732	0.4774	0.4880	0.4756	0.4726	0.4904

First, we compare the performance of HMM_{RM} and HMM. In both approaches, the weighting parameter λ was set at 0.05 and the sentence's prior distribution was assumed to be uniform. From the second and third columns of Tables 1 and 2, we observe that HMM_{RM} outperforms HMM under low summarization ratios (10% and 20%). The results demonstrate that when the original sentence model $P(w|S_i)$ in the sentence HMM was linearly combined with the relevance model $P(w|R_{S_i})$, which was constructed using information relevant to the spoken sentence, the summarization performance was improved substantially.

Next, we used the relevant information to train the sentence-dependent weighting parameter λ of the sentence HMM using Eq. (8) by the EM algorithm with ten training iterations. The fourth columns (HMM_T) in Tables 1 and 2 show the results on the development set and the evaluation set, respectively. Clearly, the summarization accuracy of HMM is significantly improved when the parameter λ is trained using the relevant information, which demonstrates the importance of estimating the HMM's parameter correctly and the efficacy of using relevant information in this way.

5.2. Estimation of the Sentence's Prior Probability Using Relevant Information

Since the importance (or prior probability) of the sentences of the spoken document to be summarized should not be identical, we tried to model the sentence's prior probability by calculating the average similarity of documents in the relevant set, i.e., we used Equations (9) and (10), where $\sum_{S_i \in D} P(S_i) = 1$, to calculate the sentence's prior probability. The fifth columns (HMM-P) in Tables 1 and 2 show the summarization results derived by HMM with a non-uniform sentence's prior probability estimated using relevant information. Comparing with the second columns (HMM) in Tables 1 and 2, we observe that HMM-P significantly outperforms HMM under low summarization ratios (10% and 20%).

The last two columns (HMM_{RM}-P and HMM_T-P) in Tables 1 and 2 show the summarization results derived by HMM_{RM} and HMM_T, respectively, where the sentence's prior probability was estimated using relevant information. It is obvious that the performance of the three approaches (HMM, HMM_{RM}, and HMM_T) with a non-uniform sentence's prior probability is significantly better than that obtained with a uniform sentence's prior probability, especially under low summarization ratios (10% and 20%). The results demonstrate that the average similarity of documents in the relevant set is a useful feature for modeling a sentence's prior probability. From Tables 1 and 2, we observe that HMM-P achieves the best summarization accuracy. The reasons why HMM_{RM}-P and HMM_T-P perform worse than HMM-P though HMM_{RM} and HMM_T outperform HMM are still under investigation.

6. CONCLUSION

We have investigated the use of information from relevant documents retrieved from a contemporary text collection for each sentence of a spoken document to be summarized in a HMM-based probabilistic generative framework for extractive spoken document summarization. The retrieved relevant text documents are used to estimate the HMM's parameters and the sentence's prior probability for each spoken sentence that consists of only a few terms or incorrect recognition terms. Very promising and encouraging results were obtained on the broadcast news summarization task.

7. ACKNOWLEDGMENTS

This work was supported in part by the National Science Council, Taiwan, under Grants: NSC95-2221-E-003-014-MY3 and NSC95-2422-H-001-031. The authors also would like to thank the NTU Speech Processing Lab for providing

the necessary speech and language data.

8. REFERENCES

[1] L. S. Lee and B. Chen, "Spoken Document Understanding and Organization," *IEEE Signal Processing Magazine*, 22(5), pp. 42-60, 2005.

[2] P. B. Baxendale, "Machine-Made Index for Technical Literature-An Experiment," *IBM Journal*, pp. 354-361, October 1958.

[3] M. Hirohata, Y. Shinnaka, K. Iwano, and S. Furui, "Sentence Extraction-Based Presentation Summarization Techniques and Evaluation Metrics," in *Proc. ICASSP 2005*.

[4] Y. Ho, "An Initial Study on Automatic Summarization of Chinese Spoken Documents", Master Thesis, National Taiwan University, July 2003.

[5] Y. Gong and X. Liu, "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis," in *Proc. ACM SIGIR 2001*.

[6] G. Murray, S. Renals, and J. Carletta, "Extractive Summarization of Meeting Recordings," in *Proc. Eurospeech 2005*.

[7] S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori, "Speech-to-Text and Speech-to-Speech Summarization of Spontaneous Speech," *IEEE Trans. Speech and Audio Processing*, 12(4), pp. 401-408, 2004.

[8] S. Maskey and J. Hirschberg, "Comparing Lexical, Acoustic/Prosodic, Structural and Discourse Features for Speech Summarization," in *Proc. Eurospeech 2005*.

[9] X. Zhu and G. Penn, "Evaluation of Sentence Selection for Speech Summarization," in *Proc RANLP 2005*.

[10] Jian Zhang and Pascale Fung, "Speech Summarization Without Lexical Features for Mandarin Broadcast News", *Proceedings of NAACL HLT 2007, Companion Volume*, pages 213-216.

[11] Y. T. Chen, S. Yu, H. M. Wang, and B. Chen, "Extractive Chinese Spoken Document Summarization Using Probabilistic Ranking Models," in *Proc. ISCSLP 2006*.

[12] B. Chen, Y. M. Yeh, Y. M. Huang, and Y. T. Chen, "Chinese Spoken Document Summarization Using Probabilistic Latent Topical Information," in *Proc. ICASSP 2006*.

[13] B. Chen and Y. T. Chen, "Word Topical Mixture Models for Extractive Spoken Document Summarization," in *Proc. ICME 2007*.

[14] Y. T. Chen, H. S. Chiu, H. M. Wang, B. Chen, "A Unified Probabilistic Generative Framework for Extractive Spoken Document Summarization," in *Proc. Interspeech 2007 – Eurospeech*.

[15] J. Xu, and W. B. Croft, "Query Expansion Using Local and Global Document Analysis," in *Proc. SIGIR 1996*.

[16] W. B. Croft, and J. Lafferty (Eds.). Language Modeling for Information Retrieval. Kluwer-Academic Publishers (2003)

[17] M. D. Smucker, D. Kulp, and J. Allan, "Dirichlet Mixtures for Query Estimation in Information Retrieval," CIIR Technical Report, Center for Intelligent Information Retrieval, University of Massachusetts (2005)

[18] B. Chen, H. M. Wang, and L. S. Lee, "A Discriminative HMM/N-Gram-Based Retrieval Approach for Mandarin Spoken Documents," *ACM Transactions on Asian Language Information Processing*, 3(2), pp. 128-145, 2004.

[19] Central News Agency (CNA) http://210.69.89.224/search/hypage.cgi?HYPAGE=login.htm

[20] C. Y. Lin, "ROUGE: Recall-oriented Understudy for Gisting Evaluation," 2003, http://www.isi.edu/~cyl/ROUGE/.