IMPROVED SPOKEN DOCUMENT SUMMARIZATION USING PROBABILISTIC LATENT SEMANTIC ANALYSIS (PLSA)

Sheng-Yi Kong and Lin-shan Lee

Speech Lab., College of EECS, National Taiwan University, Taipei, Taiwan, Republic of China
anguso@speech.ee.ntu.edu.tw, lslee@gate.sinica.edu.tw

ABSTRACT

In this paper we propose a set of new methods exploring the topical information embedded in the spoken documents and using such information in automatic summarization of spoken documents. By introducing a set of latent topic variables, Probabilistic Latent Semantic Analysis (PLSA) is useful to find the underlying probabilistic relationships between documents and terms. Two useful measures, referred to as topic significance and term entropy in this paper, are proposed based on the PLSA modeling to determine the terms and thus sentences important for the document which can then be used to construct the summary. Experiment results for preliminary tests performed on broadcast news stories in Mandarin Chinese indicated improved performance as compared to some existing approaches.

1. INTRODUCTION

In the future network era, digital content over the network will include all the information activities for human life. Clearly, the most attractive form of network content will be in multi-media including speech information, and it is in such speech information that we usually find the subjects, topics, and concepts of the associated multi-media content. As a result, spoken documents associated with network content will become key for retrieval and browsing. In other words, network content may be indexed/retrieved and browsed not only by text, but possibly by the associated spoken documents as well [1].

When considering the above network content access environment, we need to keep in mind that unlike the written documents which are better structured with titles and paragraphs and thus easier to retrieve and browse, multi-media/spoken documents are just video/audio signals, or a very long sequence of words including errors even if automatically transcribed, for example, a 3-hour video of course lecture, a 2-hour movie, or a 1-hour news episode. They are much more difficult to retrieve and browse, because they are not easily displayed on-screen, , and also because the user cannot simply "skim through" each of them from the beginning to the end. As a result, spoken document summarization — in which a summary in text or speech form is generated automatically for each spoken document (or associated multi-media content) — becomes very important [1]. Such automatically generated summaries will be very helpful in retrieving and browsing across large quantities of multi-media/spoken document archives.

Automatic summarization of text or spoken documents have been actively investigated by many groups for long time [2]. Many approaches for automatic summarization of documents, among others, have attempted to select a number of indicative sentences or passages from the original document according to a target summarization ratio, and sequence them to form a summary. Different approaches have been used to identify sentences carrying concepts closer to the complete documents [3]. The spoken documents bring extra difficulties such as the recognition errors, problems with spontaneous speech, and lack of correct sentence or paragraph boundaries. In order to avoid the redundant or incorrect parts while selecting the important and correct information in spoken documents, multiple recognition hypotheses, confidence scores, language model scores and other forms of grammatical knowledge have been utilized [4]. In recent years, a general approach have been found to be very successful [4, 5], in which each sentence in the document, $S = t_1 t_2 \dots t_j \dots t_n$, represented as a sequence of terms t_j, is given a score:

$$I(S) = \frac{1}{n} \sum_{j=1}^{n} [\lambda_1 s(t_j) + \lambda_2 l(t_j) + \lambda_3 c(t_j) + \lambda_4 g(t_j)] + \lambda_5 b(S), \quad (1)$$

where some statistical measure $s(t_j)$ (such as TF/IDF or the like) and linguistic measure $l(t_j)$ (e.g., named entities and different parts-of-speech (POSs) are given different weights, function words not included) are defined for each term t_j. $c(t_j)$ and $g(t_j)$ are calculated from the confidence score and N-gram score for the term t_j, $b(S)$ is calculated from the grammatical structure of the sentence S, and λ_1, λ_2, λ_3, λ_4 and λ_5 are weighting parameters.

In this paper we propose a set of new methods, which explores the topical information obtained in Probabilistic Latent Semantic Analysis (PLSA) modeling of terms and

documents, and uses such information to estimate the statistical measure $s(t_j)$ in equation (1) above to identify the important sentences for the topics addressed by the documents. Very encouraging summarization results have been obtained in the preliminary tests on broadcast news stories in Mandarin Chinese.

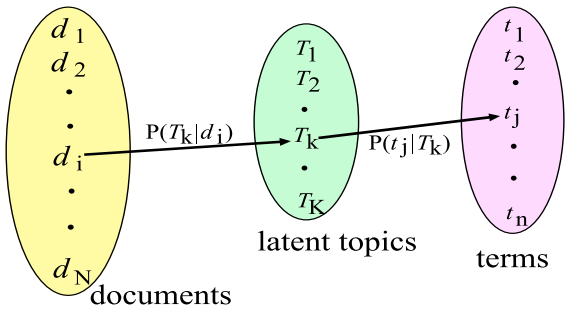

Fig. 1. Graphical representation of the Probabilistic Latent Semantic Analysis (PLSA) model.

2. BRIEF SUMMARY OF PROBABILISTIC LATENT SEMANTIC ANALYSIS (PLSA)

Latent Semantic Analysis (LSA) has been widely used in analyzing the content of documents by exploring the relationships between a set of terms and a corpus of documents considering a set of latent topics. In recent years, efforts have also been made to establish a probabilistic framework for the above latent topical approaches, including improved model training algorithms, of which Probabilistic Latent Semantic Analysis (PLSA or aspect model) [6] is often considered as a representative. In PLSA, a set of latent topic variables is defined, $T_k, k = 1, 2, \ldots, K$, to characterize the "term-document" co-occurrence relationships, as shown in Figure 1. Both the document d_i and a term t_j are assumed to be independently conditioned on an associated latent topic T_k. The conditional probability of a document d_i generating a term t_j thus can be parameterized by

$$P(t_j|d_i) = \sum_{k=1}^{K} P(t_j|T_k)P(T_k|d_i). \tag{2}$$

Notice that this probability is not obtained directly from the frequency of the term t_j occurring in d_i, but instead through $P(t_j|T_k)$, the frequency of t_j in the latent topic T_k, as well as $P(T_k|d_i)$, the likelihood that d_i addresses the latent topic T_k. The PLSA model can be optimized with the EM algorithm by maximizing a carefully defined likelihood function [6].

3. PROPOSED APPROACH

The approach proposed in this paper uses a simplified version of the widely used equation (1). We follow the successful methods reported recently [7], while focusing on the choice of the statistical measure $s(t_j)$ to be used in equation (1) using PLSA. One approach for evaluating this statistical measure $s(t_j)$ which has been proved extremely useful is called "significance score"[7](hereafter referred to as the baseline significance score),

$$s(t_j) = n(t_j, d_i) \cdot \log \frac{F_A}{F_{t_j}}, \tag{3}$$

where $n(t_j, d_i)$ is the number of occurrences of the term t_j in the given document d_i, F_{t_j} is the number of occurrences of t_j in a large corpus, and F_A is the number of occurrences of all terms or content words in the corpus. The basic idea is that terms of fewer occurrences are more semantically significant. In this paper, two new statistical measures based on PLSA modeling as summarized above are used, one based on topic significance and the other on term entropy.

3.1. Topic Significance

The topic significance score of a term t_j with respect to a topic T_k, $S_{t_j}(T_k)$, is first defined as:

$$S_{t_j}(T_k) = \sum_{d_i \in D} n(t_j, d_i) \times \frac{P(T_k|d_i)}{\sum_{T_l, l \neq k} P(T_l|d_i)}, \tag{4}$$

where $n(t_j, d_i)$ is the occurrence count of the term t_j in a document d_i, and $P(T_k|d_i)$ is obtained from a PLSA model trained with a large corpus. In equation (4) the term frequency of t_j in a document d_i, $n(t_j, d_i)$, is further weighted by a ratio which has to do with how the document d_i is focused on the topic T_k, since the denominator of the ratio is the probabilities that the document d_i is addressing all other topics different from T_k. After summation over all documents d_i, a higher $S_{t_j}(T_k)$ obtained in equation (4) implies the term t_j has a higher frequency in the latent topics T_k than other latent topics, and is thus more important in the latent topic T_k. Given this topic significance score in equation (4), the statistical measure $s(t_j)$ to be used in equation (1) based on topic significance can be defined as:

$$s_{TS}(t_j) = \sum_{k=1}^{K} S_{t_j}(T_k)P(T_k|d_i). \tag{5}$$

That is, the topic significance score of term t_j for topic T_k, $S_{t_j}(T_k)$, is further weighted by the topic distribution of the document d_i and summed over all topics. The term $P(T_k|d_i)$ can be better estimated by folding-in the probabilities $P(T_k|t_j)$. A higher $s_{TS}(t_j)$ implies the term is more important and should be given a higher priority when extracting sentences for summarization.

3.2. Term Entropy

Term entropy can be obtained from the topic distribution $P(T_k|t_j)$ of the term. We can estimate $P(T_k|t_j)$ as follows:

$$P(T_k|t_j) = \frac{P(t_j|T_k) \times P(T_k)}{P(t_j)} \approx \frac{P(t_j|T_k)}{P(t_j)}. \quad (6)$$

where the probability $P(T_k)$ is left out because a good approach to estimate it is not yet available, while $P(t_j)$ can be obtained from a large corpus. The statistical measure $s(t_j)$ to be used in equation (1) based on term entropy can then be defined as:

$$s_{EN}(t_j) = \alpha n(t_j, d_i) \left[-\sum_{k=1}^{K} P(T_k|t_j) \log P(T_k|t_j) \right], \quad (7)$$

where α is a scaling factor. Clearly, a lower entropy implies the term is more significant over the latent topics.

4. EXPERIMENTS CONFIGURATION

The preliminary experiments were performed with broadcast news stories in Mandarin Chinese. The training corpus used in the experiments included 15000 news stories in text form without word errors collected in August 2001 provided by Central News Agency of Taipei. They were used to calculate F_{t_j} and F_A in equation (3). The PLSA models were also trained using this corpus under various topic number assumptions.

The testing corpus included 200 news stories broadcast in August 2001 by a few radio stations in Taipei. The average length of each story was about 29 sec, and the speech recognition accuracy for the testing corpus for words, characters and syllables were 66.46%, 74.95% and 81.70% respectively. Each of three human subjects (students at National Taiwan University) was requested to summarize each story by ranking the importances of the sentences in each story from "most important" to "of average importance"; these summaries, three per story, were taken as references.

Sentence recall/precision has been found to be an effective evaluation metric for automatic summarization of documents [7]. Since sentence boundaries estimated by the automatic recognizer do not always agree with those in human-generated summaries, a sentence in human-generated summaries is considered to be extracted when there is an overlap of 50% or more words with the automatically generated summaries.

4.1. Special Structure of Chinese Language

The Chinese language is quite different from many Western languages, in that it has a very special structure [8]. Better use of this special structure can make spoken document processing more robust to recognition errors [9]. For the purposes here, various term definitions were chosen and used to replace the role of words, W, including characters, C, overlapping segments of two syllables, S(2), and various combinations thereof, e.g. words plus overlapping segments of two syllables, W+S(2), etc.

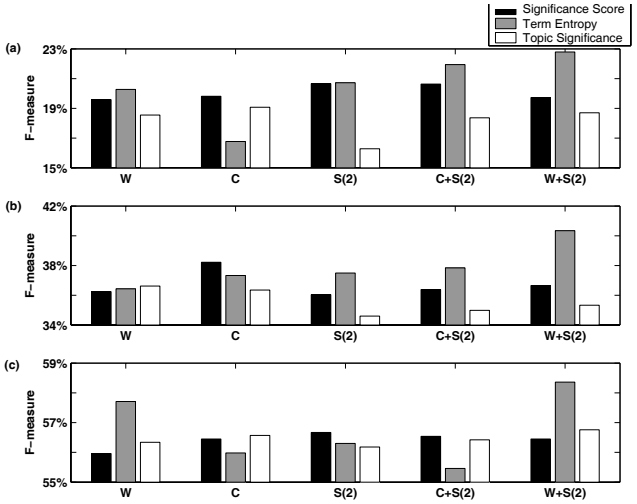

Fig. 2. F-measures obtained with the three approaches to evaluate the statistical measure in different choices of the term for (a)10%, (b)30% and (c)50% of summarization ratios.

5. EXPERIMENTAL RESULTS

The experimental results presented in terms of F-measures obtained from recall/precision rates are defined above, in Figure 2(a)(b)(c) for summarization ratios of 10%, 30% and 50% respectively. For each figure, five different term definitions were shown here, i.e., the word (W), character (C), segments of two syllables S(2), and combinations C+S(2) and W+S(2). For each case in the Figure 2, three bars were used to show the results for the three different ways to evaluate the statistical measure $s_(t_j)$ discussed in the paper: the well known, very successful significance score, the proposed term entropy, and the proposed topic significance respectively.

First consider the case of using the conventional term definition, i.e., words; the results are the first set of 3 bars in Figures 2(a)(b)(c). It can be found that the proposed term entropy (the second bar) is always significantly better than the well known significance score (the first bar), for example 21.33% vs 19.58% or a relative improvement of 8.94% for 10% of summarization ratio in Figure 2(a).The proposed topic significance (the third bar), on the other hand, is worse than the well known significance score at the 10% of summarization ratio, and only slightly better for the 30% and 50% of summarization ratios.

Next consider the different term definitions. Characters have much better recognition accuracy but carry much less semantic information as compared to words. So when comparing the results for characters with those for words (the second sets of data vs the first sets in Figures 2(a)(b)(c), characters are sometimes better (e.g. for 30% summarization ratio and significance score, 38.22% vs 36.24%), but in most cases the results are only slightly better than words, or slightly worse. Syllables have the highest recognition accuracy, but carry the least semantic information, because every syllable on average is shared by more than ten homonym characters, and can be shared by more than 50 characters. As a result, segments of two syllables (S(2) of the third sets of data in Figures 2(a)(b)(c)) have relatively unstable performance. As a result it is sometimes better than the word (for the well known significance score at 10% and 50% of summarization ratio), but similar or worse in many other cases. The combination of the above two, character plus segment of two syllables, exhibits performance closer to word, but sometimes slightly better (e.g. for the well known significance score in all ratios of 10%, 30%, and 50%), but in many other cases slightly worse. The combination of words and two-syllable segments (W+S(2)), the last sets in Figure 2(a)(b)(c), turned out to be the best with the most stable performance. It integrated the high accuracy of syllables and semantic information of words. The improvements compared to words alone were significant in some cases (e.g. 22.79% vs 21.33% or 6.84% of relative improvement for term entropy for the 10% of summarization ratio, similarly for 30% as in Figure 2(a) and (b)), and reasonably good or only slightly worse in most other cases.

Having summarized above the results of different term definitions, it is now straightforward to compare the different statistical measures discussed here in the paper. The well known significance score performed very well and stably in all cases, and it has been shown that the performance can be further improved with different term definitions other than just word (e.g. 20.66% using S(2) versus 19.58% when using W for the 10% summarization ratio, and 38.22% using C versus 36.24% when using W for the 30% summarization ratio). The topic significance proposed here is in general worse than the other two, although relatively stable performance was achieved for the 50% summarization ratio. The term entropy proposed here, on the other hand, offered the best results in most cases, and very often performed significantly better than the well known significance score: for example, for almost all of the different term definitions for the 10% and 30% summarization ratios.

Considering the above discussions, it becomes natural that at least at this moment for the given corpus of broadcast news in Mandarin Chinese, the best technique to evaluate the statistical measure is use the term entropy in equation (7) evaluated with the combinations of words and segments of two syllables. The results as compared to the well known significance score evaluated on words alone are listed in table 1. The relative improvements were significant in all summarization ratios of 10%, 30%, and 50%.

Table 1. The best results obtained in this paper (term entropy on W+S(2) as compared to the baseline well known significance score on W).

Summarization ratio	Baseline: Significance score, W	Best results: Term entropy, W+S(2)	Relative Improvement
10%	19.58%	22.79%	16.39%
30%	36.24%	40.34%	11.31%
50%	55.96%	58.36%	4.29%

6. CONCLUSION

We proposed two new methods exploring the topical information embedded in the spoken documents. Using such information, automatic speech summarization can be approved under certain conditions. Term entropy is a good measure in many cases but can still be tuned to achieve a better result.

7. REFERENCES

[1] L. S. Lee and B. Chen, "Spoken document understanding and organization," *in Special Section, IEEE Signal Processing Megazine*, 2005.

[2] I. Mani and M.T. Maybury, *Advances in Automatic Text Summarization*, Cambridge, MA:MIT Press, 1999.

[3] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proc. ACM SIGIR Conference on R&D in Information Retrieva*, 2001, pp. 19–25.

[4] J. Goldstein, M. Kantrowitz, and J. Carbonell, "Summarizing text documents: Sentence selection and evaluation metrics," in *Proc. ACM SIGIR Conference on R&D in Information Retrieva*, 1999, pp. 121–128.

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

[6] T. HofMann, "Probabilistic latent semantic analysis," *Uncertainty in Artificial Intelligence*, 1999.

[7] M. Hirohata, Y. Shinnaka, K. Iwano, and S. Furui, "Sentence extraction-based presentation summarization techniques and evaluation metrics," in *Proc. ICASSP*, 2005, pp. SP–P16.14.

[8] L. S. Lee, "Voice dictation of mandarin chinese," *IEEE Signal Processing Magazine*, vol. 14, no. 4, pp. 63–101, 1997.

[9] L. S. Lee, Y. Ho, J. F. Chen, and S. C. Chen, "Why is the special structure of the language important for chinese spoken langage processing? -examples on spoken document retrieval, segmentation and summarization," in *Proc. EUROSPEECH*, 2003, pp. 49–52.