ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken
Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

SPOKEN DOCUMENT SUMMARIZATION USING TOPIC-RELATED CORPUS AND SEMANTIC DEPENDENCY GRAMMAR

Chia-Hsin Hsieh, Chien-Lin Huang and Chung-Hsien Wu

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan, R.O.C.
{ ngsnail, chicco, chwu}@csie.ncku.edu.tw

ABSTRACT

This study presents a spoken document summarization scheme using a topic-related corpus and semantic dependency grammars. The summarization score considers speech recognition confidence, word significance, word trigram, semantic dependency grammar (SDG) and probabilistic context free grammar (PCFG). In addition, a topic-related corpus consisting of keywords as well as article is used to estimate the word significance score using latent semantic indexing (LSI). Semantic relations between words are determined by SDG using HowNet and Sinica Treebank. The dynamic programming algorithm is applied to decide the summarization ratio and look for the best summarization result according to summarization scores. Experimental results indicate that the proposed approach effectively extracts important words with semantic dependency and gives a promising speech summary.

1. INTRODUCTION

Speech is an effective way to human communication. Speech summarization is used to compress speech sources into a simple and meaningful representation. In hand-held device applications, it is appropriate to use concise summarized spoken documents instead of original multimedia documents that can save not only the transmission time over Internet but also the browsing time of users. There are other applications, such as conference or news summarization, the information index for database management and closed caption generation.

Recent summarization approaches mostly focused on text and speech topics according to different source documents. The text document summarization [1] deals with a lot of paragraphs and sentences such as articles in newspaper or literary words. The major task is to analyze the context, structure or discourse relation between paragraphs and sentences in an article, then extract important sentences to form a summary. The speech summarization [2] relies on the recognized results of a large-vocabulary continuous-speech recognizer (LVCSR) to analyze semantic and syntactic information. The difference between text and speech summarization is that speech recognition error and prosodic information should be considered in speech summarization.

In this study, we focus on spoken document summarization on TV news broadcast. There are two issues considered for speech summarization. First, a topic-related corpus is collected to choose the related keywords. Second, grammatical and semantic dependency is important for summarizing a comprehensible and meaningful sentence.

2. AUTOMATIC SPEECH SUMMARIZATION

An original speech utterance with N words is transcribed into a word sequence $X = (w_1, w_2, ..., w_N)$ using the LVCSR. Figure 1 illustrates the procedure for summarizing a spoken document. There are four steps in the summarization process:

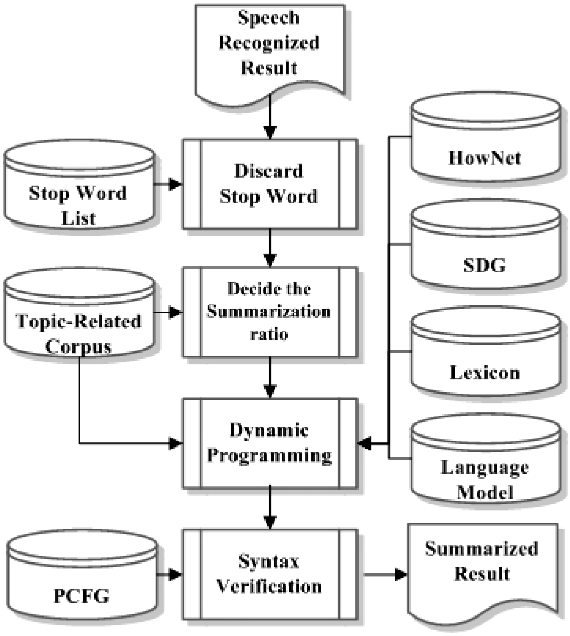

Fig. 1: Framework of spoken document summarization

First, stop words are discarded. Second, a topic-related corpus is used to extract important keywords. The average word significant score is used to estimate the compression ratio ∂. The range of ∂ is from 20% to 80% usually. In the next step, a summarized sentence $Y=(w_1,w_2,...,w_M)$ with $M=N\times\partial\times 100$ words which maximizes the following summarization score is obtained:

$$\begin{aligned}S(Y)=&\sum_{m=1}^{M}\{\lambda_C C(w_m)+\lambda_R R(w_m)\\&+\lambda_L L(w_m \mid w_{m-2},w_{m-1})+\lambda_B B_{SDG}(w_{m-1},w_m)\}\end{aligned} \qquad (1)$$

where $C(w_m)$ denotes the confidence score of word w_m obtained from the LVCSR. $R(w_m)$ denotes the word significance score. $L(w_m \mid w_{m-2},w_{m-1})$ represents the trigram probability and $B_{SDG}(w_{m-1},w_m)$ is the word concatenation score which is obtained by the semantic dependency grammar (SDG). The semantic dependency grammar and the trigram language model are used to obtain a sentence with grammatical and semantic dependency. The parameters λ_C, λ_R, λ_L and λ_B are used to balance the effect between these scores. The range of these scores X_{score} will be normalized from 0 to 1. Finally, in order to conform to the syntax, we will compute the PCFG of possible word combinations. A dynamic programming (DP) is applied to seek the best summarization result according to the summarization score.

2.1. Speech Recognition Confidence

The confidence score is used to evaluate the confidence of the recognition result. The confidence score is a posterior probability of each transcribed word, namely the ratio of a word hypothesis probability to that of all other hypothesis. It is calculated using a word graph obtained by a decoder and used as a confidence measure [3].

2.2. Word Significance

The word significance score is to evaluate the importance of words in an utterance. This study introduces a topic-related corpus with an article and its corresponding title. We collected internet news from November 2001 to September 2002 consisting of 2006 reports. First, the corresponding title words a_{ij} are extracted from each article and weighted by $tf\times idf$ [4]. Latent semantic analysis (LSA) and Singular value decomposition (SVD) [4] are adopted to reduce the noisy information. Also, a ratio r is used to keep the most important title words related to every article. In this study, r is set to 30% heuristically.

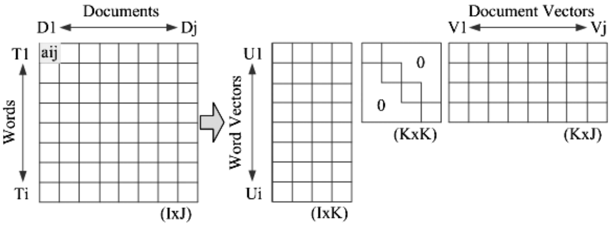

Fig. 2: Singular Value Decomposition

Therefore, these words in the title have the most important information related to the article. For the transcribed sentence X, the latent semantic indexing (LSI) is also used to retrieve a relevant document from the topic-related corpus. There are two stages to calculate the word significant score:

In the indexing stage, each document is converted into two vectors v_d^w and v_d^s, where $v_d^w=(t_d^{w_1},t_d^{w_2},...,t_d^{w_P})$ and $v_d^s=(t_d^{s_1},t_d^{s_2},...,t_d^{s_Q})$. P is the vocabulary size and Q is the number of words with syllable representation, respectively. Each $t_d^{w_m}$ is the weight of word w_m in document d and is calculated as follows:

$$t_d^{w_m}=C(w_m)\cdot[1+\ln(f_{w_m})]\cdot\ln(N/df_{w_m}) \qquad (3)$$

where $C(w_m)$ denotes the confidence of word w_m. The value is always set to 1 for indexing the topic-related text corpus, but it will vary for the query because the query is a text transcription of the speech utterance. f_{w_m} is the term frequency of word w_m in the document. df_{w_m} represents the document frequency of word w_m and N denotes the number of documents in the topic-related corpus. After the indexing stage, we obtain a word term by document matrix $A_{I\times J}^w$, shown in Fig. 2. LSI is applied using SVD to obtain $A_{I\times J}^w=U_{i\times k}^w S_{k\times k}^w V_{k\times j}^w$, where $k=\min(i,j)$ .The syllable term by document matrix also follows the same procedure to obtain $A_{I\times J}^s=U_{i\times k}^s S_{k\times k}^s V_{k\times j}^s$. After dimensionality reduction, both ${A_{I\times J}^w}^*$ and ${A_{I\times J}^s}^*$ will be used in document retrieval. Furthermore, ${A_{I\times J}^w}^*$ will be used to measure the word similarity between w_{m1} and w_{m2} as follows:

$$\begin{aligned}P_{LSI}(w_a,w_b)&=\cos(U_a^w S^w,U_b^w S^w)\\&=U_a^w S^{w2} U_b^{wT}/\left\|U_a^w S^w\right\|\left\|U_b^w S^w\right\|\end{aligned} \qquad (4)$$

In the retrieval stage, the transcribed sentence X is used as the query and converted into two vectors v_q^w and v_q^s. The most relevant document d^* is therefore retrieved as follows:

$$d^* = \arg\max_d R(q,d) \tag{5}$$

where

$$\begin{aligned} &R(q,d)\\ &= \alpha_R \cos(v_q^w S^w, v_d^w S^w) + (1-\alpha_R)\cos(v_q^s S^s, v_d^s S^s)\\ &= \alpha_R \cdot (v_q^w S^{w2} v_d^{wT}) / (\|v_q^w S^w\| \cdot \|v_d^w S^w\|)\\ &+(1-\alpha_R)(v_q^s S^{s2} v_d^{sT}) / (\|v_q^s S^s\| \cdot \|v_d^s S^s\|) \end{aligned} \tag{6}$$

Parameter α_R is used to balance the effect of word and syllable feature in the retrieval module. After retrieving the most relevant document d^* , the words in the corresponding title t^* contain the most important information related to d^* . We calculate the word significance score $R(w_a)$ of w_a in the transcribed sentence according to the words in title t^* and the word co-relation matrix which was obtained from LSI:

$$R(w_a) = \max_b \{P_{LSI}(w_a, w_b^{t^*}) \cdot f_{w_a} \cdot \ln(N / df_{w_a})\} \tag{7}$$

where $P_{LSI}(w_a, w_b^{t^*})$ denotes the similarity between word w_a and $w_b^{t^*}$. Following this procedure, we can extract the related important information from the topic-related corpus. Nevertheless, if there is no related document for query q , the word significance score will be replaced by its original term weighting as in Eq. (3).

2.3. Word Trigram

The word trigram score L(w_m|$w_{m\text{-}1}$,$w_{m\text{-}2}$) is used to evaluate the probability of a summarized sequence.

2.4. Semantic Dependency Grammar

This study proposes a modified semantic dependency grammar (SDG) [4] to obtain the semantic concatenation score $B_{SDG}(w_a, w_b)$ as follows:

$$\begin{aligned} &B_{SDG}(w_a, w_b)\\ &= \frac{1}{N_s} \sum_{j=1}^{N_s} \sum_i \sum_r f_{DR_i^r(w_a, w_b)}(T_i, S^j(w_a, w_b)) \times f_{T_i}(S^j(w_a, w_b)) \end{aligned} \tag{8}$$

where $S^j(w_a, w_b)$ denotes sentence S^j containing words w_a and w_b . $f_{T_i}(.)$ denotes the score of PCFG . T_i denotes the parsing tree. $f_{DR_i^r}(.)$ denotes the score of SDG. N_s denotes total sentence numbers. Dependency graph $D_i = \{DR_i^r(w_a, w_b) \,|\, 1 \le r \le N_w - 1\}$ denotes the set of dependency relation DR_i^r from the parsing tree T_i and sentence $S^j(w_a, w_b)$ consisting of N_w words.

In order to avoid the sparse data problem when estimating dependency relations, we use the PCFG of Sinica Treebank [6] to parse the sentence and replace each word with its own Hypernym based on HowNet.

$$f_{DR_i^r(w_a, w_b)}(T_i, S^j(w_a, w_b)) \cong f_{DR_i^r(H(w_a), H(w_b))}(T_i, S^j(w_a, w_b)) \tag{9}$$

where $H(w_a)$ denotes the Hypernym of w_a, Furthermore, the score is estimated using following equations:

$$\begin{aligned} &f_{DR_i^r(H(w_a), H(w_b))}(T_i, S^j(w_a, w_b))\\ &= C(DR_i^r(H(w_a), H(w_b))) / C(H(w_a), H(w_b)) \end{aligned} \tag{10}$$

where $C(DR_i^r, H(w_a), H(w_b))$ denotes the frequency that dependency relation $DR_i^r(H(w_a), H(w_b))$ happens in training corpus. $C(H(w_a), H(w_b))$ denotes the co-occurrences of $H(w_a)$ and $H(w_b)$ in the training corpus

Figure 3 illustrates an example of a dependency graph which is constructed from the sentence“ 我們(We) 遊覽(go sightseeing) 台灣(Taiwan) 各個(every) 景點(scenic spot).”

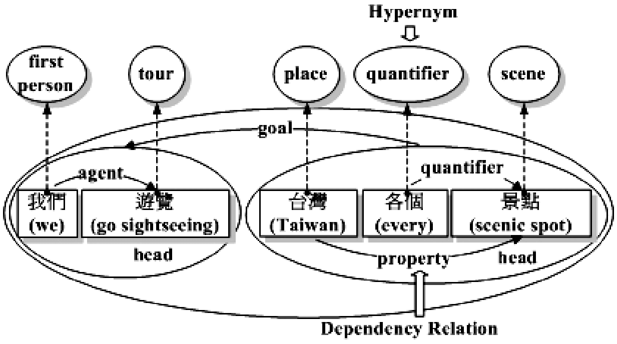

Fig. 3: Example of dependency graph

3. EXPERIMENTS

This study conducted two experiments to evaluate the performance of the proposed framework. The first experiment is an objective evaluation which evaluates the ability of key information extraction using information retrieval (IR) technique. The second experiment is a subjective evaluation which evaluates the grammatical and semantic dependency of the summarized result.

An HMM-based Mandarin LVCSR was constructed. In the speech recognizer, spectral analysis is conducted using a 16 ms frame shifted by 8 ms. For each frame, 12 me-frequency cepstral coefficients (MFCC) and the logarithmic energy are extracted, and these coefficients are combined with their first and second time derivatives to form a 39-dimensional feature vector. The training data of acoustic model consisting of 4 hours anchor speech from TV news from 2001 to 2002. The language model was trained using a newswire text corpus from News website consisting of 20 million Chinese characters in the same time. The character accuracy is 82%.

Furthermore, the semantic dependency grammar (SDG) was constructed from the Sinica Treebank [6], Taiwan, with 36953 sentences and HowNet [7]. We extracted 22,025 rule sets according to the tree structure of Part-of-Speeches (POSs) and obtain their corresponding probabilities from the Treebank. The evaluation data for speech summarization is collected from TV news broadcast in 2002. There are 100 spoken documents consisting of 845 utterances by a female anchor speaker. Furthermore, the parameters λ_C , λ_R , λ_L and λ_B are empirically set to 0.1, 0.2, 0.4 and 0.3 respectively. We conducted two experiments to evaluate the performance:

3.1. Evaluation of Key Information Extraction

The first experiment is an objective evaluation which evaluates the ability of key information extraction using information retrieval (IR) technique. In this experiment, we used a task-based approach to evaluate the ability for key information extraction. Twenty keyword queries were assigned to retrieve both the transcribed evaluation data (100 stories) and summarized results. For each query, the system returned 10 retrieved documents. Two performance measures were used: the mean average precision (mAP) and the raw average precision (rAP) [8]. The mAP and rAP are calculated as follows:

$$\text{mAP} = \frac{1}{N_q}\sum_{i=1}^{N_q}\frac{1}{N_i}\sum_{k=1}^{N_i}\frac{k}{rank_{ik}} \; ; \; \text{rAP} = \frac{1}{N_q}\sum_{i=1}^{N_q}\frac{N_i}{N} \qquad (11)$$

where N_q denotes the number of queries, N_i denotes the number of relevant documents contained in the retrieved documents for query q_i and $rank_{ik}$ denotes the rank of the kth relevant document for query q_i .

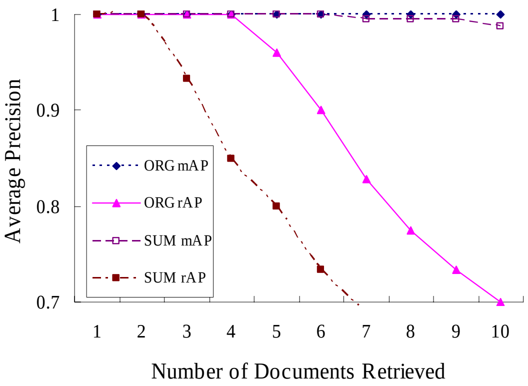

Fig.4: Experimental results for key information extraction

In Figure 4, high mAP indicates that the retrieved relevant documents appear in higher ranks among the retrieved documents. Low rAP is obtained when the number of retrieved documents increases because the evaluation database only consists of 100 stories and some queries only have a few relevant documents. We can observe that summarized results achieve the same mAPs as original transcriptions. In other words, the redundant information is not much.

3.2. Evaluation of Word Accuracy Compared with Manual Summarization Results

The second experiment evaluates the word accuracy compared with manual summarization results. Two graduate students were asked to summarize news articles and the word accuracy is obtained by following formula:

$$P_{accuracy} = (W - I - D - S)/W \qquad (12)$$

where W is the word length. I denotes the insertion errors. D denotes the deletion errors and S is the substitution errors. Table I shows the experimental results. We can observe that there is a high insertion error rate due to the subjective variation between different people and the proposed approach.

Table I. Experimental results of word accuracy

	Accuracy	Insertion	Deletion	Substitution
Percentage (%)	35.05	32.1	13.8	21.1

4. CONCLUSION

This study has presented a topic-related corpus approach and a semantic dependency grammar for spoken document summarization. A topic-related corpus was used to extract important information. A dynamic programming technique was used to search the best summarization result efficiently. The experimental results demonstrate that the proposed framework achieves a satisfactory performance.

5. REFERENCES

[1] I. Manu and M. Maubury, *Advances in Automatic Summarization*. Cambridge, MA: MIT Press, 1999.

[2] C. Hori and S. Furui, “A new approach to automatic speech summarization,” *IEEE Trans. on Multimedia*, vol. 5, no. 3, pp. 368-378, 2003.

[3] T. Kemp and T. Schaaf, “Estimating confidence using word lattices, “ *in Proc. 5th Eurospeech*, vol. 2, Rhodes, Greece, 1997, pp. 827-830.

[4] Christopher D. Manning and Hinrich Schutze, “Foundations of Statistical Natural Language Processing,” The MIT Press, 1999

[5] Lawrence Rabiner , Biing-Hwang Juang, “Fundamentals of speech recognition,” Prentice-Hall, Inc., Upper Saddle River, NJ, 1993

[6] CKIP Treebank http://godel.iis.sinica.edu.tw/CKIP/treebank/

[7] HowNet, http://www.keenage.com/

[8] M. Banko, V. Mittal and M. Witbrock, “Headline generation based on statistical translation, “ *in Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 2000, pp. 318-325.