AUTOMATIC TITLE GENERATION FOR CHINESE SPOKEN DOCUMENTS WITH A DELICATE SCORED VITERBI ALGORITHM

Sheng-yi Kong, Chien-chi Wang, Ko-chien Kuo, 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

Automatic title generation for spoken documents is believed to be an important key for browsing and navigation over huge quantities of multimedia content. A new framework of automatic title generation for Chinese spoken documents is proposed in this paper using a delicate scored Viterbi algorithm performed over automatically generated text summaries of the testing spoken documents. The Viterbi beam search is guided by a delicate score evaluated from three sets of models: term selection model tells the most suitable terms to be included in the title, term ordering model gives the best ordering of the terms to make the title readable, and title length model tells the reasonable length of the title. The models are trained from a training corpus which is not required to be matched with the testing spoken documents. Both objective evaluation based on F1 measure and subjective human evaluation for relevance and readability indicated the approach is very attractive.

Index Terms— Spoken documents, title generation

1. INTRODUCTION

With the ever-increasing bandwidth of Internet and fast falling memory costs, multimedia and audio information has become a very important part of network content, such as broadcast programs, lecture and meeting records, as well as many other video/audio documents. However, multimedia and audio content are very difficult to be shown on the screen, and thus very difficult to browse. Automatic title generation is believed to be a very important key for easy browsing and navigation over such content. It will be much more easier if each multimedia document (or its segment) can be given a title in text automatically based on the included speech information.

A title is different from a summary, in addition to being especially short. It is also different from a set of key words. It needs to express the concepts carried by the entire documents in a sequence of only several words or phrases in a readable form. This is why automatic title generation is challenging.

A non-extractive statistical model for title generation was first proposed [1], which can generate titles for documents by statistically learning from documents and corresponding human-generated titles. This model was then extended by some more advanced techniques later on [2, 3]. More interesting approaches were also developed [4, 5]. A very efficient Adaptative K Nearest-Neighbor approach (AKNN) for title generation was also proposed [6], which was based on the assumption of the availability of a training text corpus with human-generated titles which covers the subject areas of the testing spoken documents (or matched training corpus). The basic idea of AKNN is to select a document in the training corpus most close to the testing spoken document, and use its human-generated title as the title for the testing spoken document, with a few key terms possibly replaced by the key terms in the testing spoken document. This approach offers highly readable titles since they are actually human-generated, but requires a well matched training corpus. If the training corpus is not matched, the automatically generated title may be completely irrelevant to the testing spoken document.

In this paper we propose a new framework of automatic title generation using a delicate scored Viterbi algorithm performed over automatically generated text summaries of the testing spoken documents. The scores used in the Viterbi algorithm are obtained by three sets of carefully designed models trained from a training corpus: term selection models, term ordering model, and title length model. A training corpus with human-generated titles are still needed, but not required to be matched to the subject areas of the testing spoken documents.

2. PROPOSED APPROACH

2.1. Overview Of The Proposed Approach

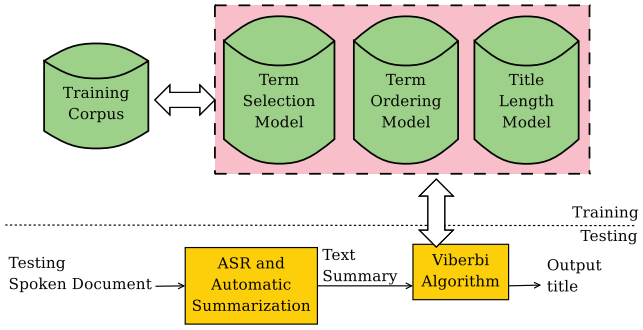

Fig. 1. Overview of the proposed approach.

The framework for the proposed approach includes two parts: the training part and the testing part, as shown in Fig. 1. In the training part, three sets of carefully designed models are trained with a training corpus of text documents with human-generated titles: term selection model, term ordering model and title length model. In the testing part, the input testing documents are first transcribed into Chinese texts with errors using ASR techniques, and then text summaries [7] are obtained. In this way the least important utterances can be removed and important terms can be better collected and used to construct the title. A delicate Viterbi algorithm is then performed on the summaries with scores obtained from the above three sets of models. This gives the output title. Below we assume the training corpus D includes N text documents, $D = \{d_1, d_2, \ldots, d_N\}$, with

a corresponding human-generated title set $T = \{t_1, t_2, \ldots, t_N\}$, where t_i is the human-generated title of d_i.

2.2. Term Selection Model

Those terms often used in titles are referred to as "title term" in this paper. There are three types of title terms: named entities, key terms which are not named entities (e.g. earthquake), and general terms often used in titles which are not key terms (e.g. announce, today). All these three types of terms should be selected. To select proper terms w_j to be used in the title for a testing spoken document $\bar{d}_i$, a single term selection score $S_1(w_j, \bar{d}_i)$, obtained from the term selection model is used, which is the weighted sum of six scores,

$$\begin{aligned} S_1(w_j, \bar{d}_i) = c_1 H_{\text{NE}}(w_j) + \\ \log\left[\mathrm{H}_{\text{EN}}(w_j)^{c_2} \cdot \mathrm{H}_{\text{title}}(w_j)^{c_3} \cdot \mathrm{G}(w_j, \bar{d}_i)^{c_4}\right] + \\ c_5 \cdot \mathrm{R}(w_j, \bar{d}_i) + c_6 \cdot \Omega(w_j, \bar{d}_i), \end{aligned} \quad (1)$$

where each score is respectively explained below.

2.2.1. Named Entity Score

Named entities are always important in a spoken document. They need to be extracted, and given a score $H_{\text{NE}}(w_j)$ to indicate their importance. $H_{\text{NE}}(w_j)$ is a constant if w_j is a named entity and zero otherwise.

2.2.2. Latent Topic Entropy Score

It has been found that the "Latent Topic Entropy" derived from Probabilistic Latent Semantic Analysis (PLSA) [8] is a very useful measure to identify key terms from a document [9]. Latent Topic Entropy $\mathrm{EN}(w_j)$ for a term w_j is defined as

$$\mathrm{EN}(w_j) = -\sum_{k=1}^{K} P(z_k|w_j) log[P(z_k|w_j)], \quad (2)$$

where $\{z_k, k = 1, 2, \ldots, K\}$ is the set of latent topics derived from PLSA, $P(z_k|w_j)$ is the probability of a latent topic z_k being discussed when observing the term w_j, which can be obtained from the PLSA model. So when w_j is important to only a few latent topics, the distribution of $P(z_k|w_j)$ over all z_k is usually more focused over these latent topics, and the Latent Topic Entropy $\mathrm{EN}(w_j)$ is lower. Thus terms with lower Latent Topic Entropy is more likely to be a key term. Many key terms can be identified in this way, including named entities and those which are not named entities. Therefore, we define the Latent Topic Entropy score $\mathrm{H}_{\text{EN}}(w_j)$ as:

$$\mathrm{H}_{\text{EN}}(w_j) = \frac{1}{1 + \mathrm{EN}(w_j)^h}, \quad (3)$$

where h is a weighting factor and for smoothing purpose we add 1 in the denominator.

2.2.3. Title Term Probability Score

The title term probability score $\mathrm{H}_{\text{title}}(w_j)$ for a term w_j is proportional to the probability of the term w_j to appear in the title t_i of a document d_i given w_j is observed in d_i. It is obtained by:

$$\mathrm{H}_{\text{title}}(w_j) = \frac{\sum_{t_k \in T} I(w_j, t_k)}{\mathrm{df}(w_j)}, \quad (4)$$

where

$$I(w_j, t_k) = \begin{cases} 1, \text{if } w_j \text{ appears in title } t_k, \\ 0, \text{otherwise}, \end{cases} \quad (5)$$

and $\mathrm{df}(w_j)$ is the total number of documents d_i which includes the term w_j. This score is useful to identify all title terms including those which are not key terms.

2.2.4. NBL Score

Naïve Bayesian Approach with Limited Vocabulary (NBL) score $\mathrm{G}(w_j, \bar{d}_i)$ was derived to measure the potential of a term w_j in the testing document $\bar{d}_i$ to be used in the corresponding title $\bar{t}_i$[6], as defined by

$$\mathrm{G}(w_j, \bar{d}_i) = \mathrm{tf}(w_j, \bar{d}_i) \times P(w_j, T|w_j, D), \quad (6)$$

where $\mathrm{tf}(w_j, d_i)$ is the frequency count of the term w_j in the document d_i, and

$$P(w_j, T|w_j, D) = \frac{\sum_{i=1}^{N} \mathrm{tf}(w_j, t_i) \times \mathrm{tf}(w_j, d_i)}{\sum_{i=1}^{N} \mathrm{tf}(w_j, d_i)}, \quad (7)$$

where $\mathrm{tf}(w_j, t_i)$ is the frequency count of the term w_j in the training title t_i for the training document d_i. This score is also useful to identify all title terms which are not key terms.

2.2.5. Rank Score

Ranking is usually more stable than raw scores. We therefore rank all terms w_j in the testing document $\bar{d}_i$ using the weighted sum of $\mathrm{H}_{\text{EN}}(w_j)$, $\mathrm{H}_{\text{title}}(w_j)$ and $\mathrm{G}(w_j, \bar{d}_i)$ in Eqs. (3)(4)(6) to give an integer $\mathrm{rank}(w_j, \bar{d}_i)$. The rank score $\mathrm{R}(w_j, \bar{d}_i)$ for term w_j is thus defined as:

$$\mathrm{R}(w_j, \bar{d}_i) = \left[\mathrm{rank}(w_j, \bar{d}_i)\right]^{-1}, \quad (8)$$

so those terms with higher ranks have higher scores.

2.2.6. Position Score

Very often the first several sentences in a document include the most important terms. We therefore define a position score $\Omega(w_j, \bar{d}_i)$ for a term w_j in a document $\bar{d}_i$ as:

$$\Omega(w_j, \bar{d}_i) = \left[\mathrm{st}(w_j, \bar{d}_i)\right]^{-1}, \quad (9)$$

where $\mathrm{st}(w_j, \bar{d}_i)$ is the sequence order of the sentence in the testing document $\bar{d}_i$ where the term w_j is first observed.

2.3. Title Ordering Model

With title terms properly selected, we need to order them reasonably to produce more readable titles. In the proposed method, 4 versions of tri-gram language model were used. This includes 2 types of tri-gram language model: one for terms and the other for POS (part-of-speech) tags of terms, each with 2 versions, one trained with the training document set D and the other with the corresponding training title set T only. This is because the titles are usually more condensed and more elegant, thus can be used to train different language models. The overall tri-gram language model is therefore the weighted average of the 4 versions of tri-gram language model:

$$\bar{P}(w_n|w_{n-2}, w_{n-1}) = P_1^{b_1} \cdot P_2^{b_2} \cdot P_3^{b_3} \cdot P_4^{b_4} \quad (10)$$

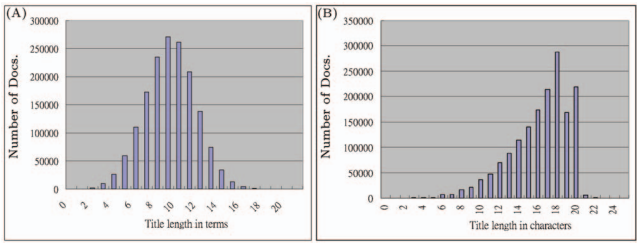

Fig. 2. Histograms of document numbers with different title length measured in number of (a)terms and (b)characters.

where P_1,P_2,P_3,P_4 are the tri-gram probabilities obtained with the 4 versions of language model, and b_1,b_2,b_3,b_4 are the respective weights. Therefore the term ordering score $S_2(\bar{t}_i)$ for a term sequence or candidate title $\bar{t}_i$, given by the term ordering model here, can be evaluated accordingly.

2.4. Title Length Model

We found that a good title should have a suitable length. Very often a too long title natural includes noisy terms, while a too short title has some important terms missing. We therefore developed a title length model from the training title set T of the training corpus, which include all human-generated titles. Chinese is not alphabetic. Each Chinese character has its own meaning. A Chinese term may consist of one to several characters. Hence the length of a title can be measured in either number of terms or number of characters.

By analyzing the title length statistics for the training corpus to be described below, it was found that the distribution of title length in number of terms is close to a Gaussian distribution, as shown in Figure 2(a). However, as can be seen in Figure 2(b), the distribution of title length in number of characters is completely different. Most of human-generated titles have 14-20 characters, which is clearly a preferred length, and very few of them exceed 20.

Considering the two measures of term number and character number, we define the title length score for a title with a length of n_w terms and n_c characters as:

$$S_3(n_w, n_c) = \log\left[P_{term}(n_w)^{g_1} \cdot P_{char}(n_c)^{g_2}\right], \text{where} \quad (11)$$

$$P_{term}(n_w) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left\{-\frac{(n_w-\mu)^2}{2\sigma^2}\right\}, \quad (12)$$

$$P_{char}(n_c) = \frac{T_{n_c}}{N}, \quad (13)$$

where μ and σ are the mean and variance of the Gaussian distribution in Fig. 2(a), T_{n_c} is the number of training titles with length equals to n_c characters, N is the total number of training titles, so Eqs. (12)(13) are simply the two distributions in Fig. 2(a)(b), and g_1 and g_2 are weighting factors.

2.5. Viterbi Beam Search

As shown in the lower part of Fig. 1, Viterbi search is performed over the automatically generated text summary for the testing spoken document, in which every term in the text summary is a state, every state can transit to every other state, while the path linking a sequence of states (terms) giving the highest score from the above three sets of models is the output title,

$$t_i^* = \arg\max_{\bar{t}_i}\Big\{\frac{\alpha}{n_w(\bar{t}_i)}\sum_{w_j\in\bar{t}_i} S_1(w_j,\bar{d}_i) + \frac{\beta}{n_w(\bar{t}_i)}S_2(\bar{t}_i) + \gamma S_3(n_w(\bar{t}_i), n_c(\bar{t}_i))\Big\}, \quad (14)$$

where $S_1(w_j,\bar{d}_i)$, $S_2(\bar{t}_i)$, $S_3(n_w(\bar{t}_i), n_c(\bar{t}_i))$ are the scores from the three sets of models mentioned above, $\bar{t}_i$ is a candidate title for the testing document $\bar{d}_i$, $n_w(\bar{t}_i)$ and $n_c(\bar{t}_i)$ are the length of $\bar{t}_i$ measured by terms and characters, α, β, γ are weighting parameters, and t_1^* is the output title. Beam search is used here to reduce the search space.

3. EXPERIMENTAL RESULTS

3.1. Experimental Setup

Broadcast news stories were taken as examples of spoken documents. The testing corpus included 118 news stories collected from radio stations in Taipei in Dec 2005. Two sets of training corpora were used, both were text news stories collected from news agencies in Taipei with human-generated titles. Training set 1 included 10,660 news stories collected in Jan 2001, and training set 2 included 7,523 news stories collected in Dec 2005. So training set 1 was mismatched while training set 2 was matched. The spoken document summarization was primarily based on PLSA [9], while the named entity recognition used various approaches including global evidences and external knowledge sources [10].

The reference titles for testing spoken documents were produced by the students of the Graduate Institute of Journalism of National Taiwan University. These reference titles were used in the objective performance measures presented below.

The objective performance measures included precision, recall and F1 scores, where precision and recall were calculated from the number of identical Chinese terms in automatically generated and human-generated titles. In addition, five-level subjective human evaluation was also performed, where 5 was the best and 1 was the worst. Two different metrics were used in the subjective human evaluation, "Relevance" calibrating the relation between the automatically generated titles and the testing spoken documents, and "Readability" indicating how the automatically generated title is readable. In performing the subjective human evaluation, each subject was given the reference titles with reference scores for both "Relevance" and "Readability" of 5, 3 and 1 for some reference documents, so the results can be more consistent for different subjects. 23 subjects participated in the test.

3.2. Experimental Results

The results of objective evaluation are listed in Table 1, where two baseline approaches were compared with the proposed approach, the non-extractive statistical model [1] as baseline 1 (BL1) and AKNN approach [6] as baseline 2 (BL2). The results for the mismatched training set 1 (TS1) are in the upper half of the table, while those for matched training set 2 (TS2) are in the lower half.

In all cases in Table 1, a matched training set gave much better results (TS2 vs TS1). BL2 (AKNN) was specially sensitive to the matched condition, as mentioned previously, so it performed reasonably well with TS2, but very poorly with TS1. But in all cases the proposed scored Viterbi not only performed significantly better than

Training sets	Approaches	F1	Precision	Recall
TS1 (mismatched)	BL1(statistical)	0.0876	0.0649	0.1344
	BL2(AKNN)	0.0420	0.0404	0.0437
	Scored Viterbi	0.1733	0.1452	0.2148
TS2 (matched)	BL1(statistical)	0.1032	0.0783	0.1514
	BL2(AKNN)	0.1183	0.1315	0.1074
	Scored Viterbi	0.1938	0.1933	0.1943

Table 1. Objective evaluation of the proposed scored Viterbi algorithm as compared to the two baselines, statistical (BL1) and AKNN (BL2). Training set 1 (TS1) was mismatched with the testing set, while Training set 2 (TS2) was matched.

Training sets	Approaches	Relevance	Readability
TS1 (mismatched)	BL1(statistical)	3.491	1.863
	BL2(AKNN)	2.463	4.611
	Scored Viterbi	4.187	3.917
TS2 (matched)	BL1(statistical)	3.615	1.874
	BL2(AKNN)	3.594	4.615
	Scored Viterbi	4.259	4.053

Table 2. Subjective human evaluation of the proposed scored Viterbi algorithm as compared to the two baselines, statistical (BL1) and AKNN (BL2). Training set 1 (TS1) was mismatched with the testing set, while Training set 2 (TS2) was matched.

the two baseline approaches, but was much more robust with respect to the matched condition of the training set.

The F1 measures look relatively low here (0.1733 or 0.1938). In a separated work three different students were asked to produce titles for another set of spoken documents, and the averaged cross-evaluated F1 measures among the three titles produced by the three different students for the same spoken document was found to be between 0.454 and 0.512. In other words, different people produce quite different titles. Compared to human-generated results (0.454 to 0.512), the numbers here (0.1733 or 0.1938) seemed to be reasonable.

The corresponding subjective human evaluation scores are listed in Table 2. BL2 (AKNN) was again very sensitive to matched conditions in the relevance score (very poor for TS1 but reasonable for TS2), although excellent for both in readability. This is reasonable considering the nature of AKNN approach. But the proposed approach performed very well for both relevance and readability with both matched and mismatched conditions.

It is interesting to find out the importance of each approach proposed here by examine the degradation of F1 score when one of the approaches was removed from the complete proposed scored Viterbi. The results with the mismatched TS1 are listed in Table 3. Case (a) is the complete proposed scored Viterbi. Cases (b)(c)(d)(e) are those when one of the scores in the term selection model is deleted. It turned out that all these scores are important, and in particular the title term probability score $\mathrm{H}_{\text{title}}(w_j)$ played very important role to produce good title terms to be used in the titles. Case (f) is for the POS tag tri-gram in the term ordering model, cases (g)(h) for title length model, and case (i) for starting with the text summary. Again all these approaches are important, and in particular the Gaussian model for title length in number of terms turned out to be a very important key.

Training sets		Approaches	F1
TS1 (mismatched)	(a)	Scored Viterbi	0.1733
	(b)	Scored Viterbi without $\mathrm{H}_{\text{EN}}(w_j)$ in Eq. (3)	0.1575
	(c)	Scored Viterbi without $\mathrm{H}_{\text{title}}(w_j)$ in Eq. (4)	0.1322
	(d)	Scored Viterbi without $\mathrm{G}(w_j, d_i)$ in Eq. (6)	0.1608
	(e)	Scored Viterbi without $\Omega(w_j, d_i)$ in Eq. (9)	0.1624
	(f)	Scored Viterbi without POS tag tri-gram	0.1609
	(g)	Scored Viterbi without P_{term} in Eq. (12)	0.1325
	(h)	Scored Viterbi without P_{char} in Eq. (13)	0.1534
	(i)	Scored Viterbi without Summarization	0.1549

Table 3. Degradation of F1 measure of the proposed scored Viterbi when one of the included approaches was deleted.

4. CONCLUSION

In this paper we proposed a delicate scored Viterbi algorithm for automatic title generation for Chinese spoken documents. The proposed framework is found to be quite effective in generating reasonably good titles, and quite robust to mismatch between training and testing sets. These were verified by both objective and subjective evaluations.

5. REFERENCES

[1] Michael J. Witbrock and Vibhu O. Mittal, "Ultra-summarization: A statistical approach to generating highly condensed non-extractive summaries," in *Proc. of ACM SIGIR*, 1999, pp. 315–316.

[2] Michele Banko, Michael J. Witbrock, and Vibhu O. Mittal, "Headline generation based on statistical translation," in *Proc. of ACL*, 2000.

[3] Stephen Wan, Mark Dras, Cecile Paris, and Robert Dale, "Using thematic information in statistical headline generation," in *Proc. of ACL*, 2003.

[4] Rong Jin and Alex G. Hauptmann, "Title generation for spoken broadcast news using a training corpus," in *Proc. of ICSLP*, 2000.

[5] R. Jin and A. Hauptmann, "Automatic title generation for spoken broadcase news," in *Proc. of HLT*, 2001, pp. 1–3.

[6] Shun-Chuan Chen and Lin-shan Lee, "Automatic title generation for chinese spoken documents using an adaptive k nearest-neighbor approach," in *Proc. in EUROSPEECH*, 2003, pp. 2813–2816.

[7] 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.

[8] Thomas. Hofmann, "Probabilistic latent semantic analysis," in *Proc. of the 15th Conference on Uncertainty in AI*, 1999.

[9] Sheng-yi Kong and Lin-shan Lee, "Improved spoken document summarization using probabilistic latent semantic analysis (plsa)," in *Proc. of ICASSP*, 2006.

[10] Yi-chen Pan, Yu-ying Liu, and Lin-shan Lee, "Named entity recognition from spoken documents using global evidence and external knowledge sources with application on mandarin chinese," in *IEEE Conf. Proc. of Automatic Speech Recognition and Understanding*, 2005, pp. 296–301.