UTTERANCE-LEVEL LATENT TOPIC TRANSITION MODELING FOR SPOKEN DOCUMENTS AND ITS APPLICATION IN AUTOMATIC SUMMARIZATION

Hung-yi Lee #1, Yun-nung Chen #2, Lin-shan Lee #3

Graduate Institute of Communication Engineering, National Taiwan University#
Graduate Institute of Computer Science and Information Engineering, National Taiwan University*
tlkagkb93901106@gmail.com¹, vivian.ynchen@gmail.com², lslee@gate.sinica.edu.tw³

ABSTRACT

In this paper, we propose to use an utterance-level latent topic transition model to estimate the latent topics behind the utterances, and test the performance of such model in extractive speech summarization. In this model, the latent topic weights behind an utterance are estimated, and these topic weights evolve from an utterance to the next in a spoken document based on a topic transition function represented by a matrix. We explore different ways of obtaining such topic transition matrices used in the model, and find using a set of matrices estimated with utterances clustered from a training spoken document set is very useful. This model was shown to be able to offer extra performance improvement when used with the popularly used Probability Latent Semantic Analysis (PLSA) in preliminary experiments on speech summarization.

Index Terms— Latent Topic Transition Modeling, Speech Summarization

1. INTRODUCTION

Latent topics have been widely used for analysing both text and spoken documents by discovering word clustering patterns in the documents and somehow projecting each document to a latent topic space constructed through such word clusters. Such latent topic information has been found very useful in many applications such as concept matching in information retrieval, document summarization and clustering, and key term extraction. In many cases, it is desired to have more precise latent topic information on sentence level (or utterance level) rather than for each document, and the latent topic information for each sentence (utterance) can be estimated by simply treating a sentence (utterance) as a short document and directly applying the latent topic analysis methods developed for documents. However, since a sentence (utterance) usually contains only several terms, very often the latent topics cannot be well estimated unless the context is considered. For example, assume a news story about golf is being analysed. An utterance with the term “Tiger Woods” but without the term “golf” may be viewed as about “animals” although it is very likely to be more related to “sports”. Therefore, utterance-level latent topic analysis may not be as straightforward as document-level. The problem is more difficult for spoken documents with serious recognition errors since it is possible that very limited terms are recognized correctly within an utterance. Moreover, utterance boundaries are not clear in a spoken document, so automatically segmented utterances may not be as well formed as in text. In reality, natural speech rarely consists of isolated, unrelated utterances but rather collocated, structured and coherent utterances. Hence, an utterance with the term “Tiger Woods” may be found to be about “sports” if the neighboring sentences are about “sports”, even if no terms related to “sports” actually appear in the utterance being considered.

Although unsupervised learning of linguistic structure have been widely studied[1], not too many topic models have ever attempted to model similar structural dependency among topics. Hidden Markov Models have been successfully used for capturing topic transition in summarization [2, 3], but the assumption that each sentence is generated from a single latent topic instead of a topic mixture may not be sufficient. The Hidden Topic Markov Model (HTMM)[4] was extended from Latent Dirichlet Allocation by modeling the topic dependency between adjacent sentences, but in HTMM the topic dependency is simply binary: the topics of a sentence are either independent of or exactly the same as the previous sentence. Structural Topical Model (strTM)[5] further improved HTMM by assuming that adjacent sentences follow a topic transition relation, but it is on text documents only rather than spoken documents.

In this paper, we propose to use an utterance-level latent topic transition model for spoken documents, in which a topic transition function is used to model the change of latent topic weights across adjacent utterances, so the topic weights of an utterance depend not only on the terms in the utterance being considered, but also on the topics of the preceding utterance. We then apply this model in extractive speech summarization, in which a number of indicative utterances was selected from the given spoken documents according to a target summarization ratio, and concatenated together to form the summary. In such speech summarization tasks, it is important to identify utterances that carry concepts closer to the document as a whole, and it has been verified that measuring the topic-based utterance-document similarity is more effective than the word-based alternative, because the former is vulnerable to problems like synonyms and recognition errors. By enhancing the topic estimation process by topic transition modelling, more accurate latent topic weights in each sentence can be used for sentence-document similarity measure.

2. SENTENCE-LEVEL LATENT TOPIC TRANSITION MODEL

The proposed utterance-level topic transition model for spoken documents is shown in Fig. 1. \( \theta_i \) is a \( K \)-dimensional topic weight vector for the \( t \)-th utterance in the spoken document, and \( K \) is the number of topics. \( \theta_i[k] \) is the \( k \)-th component of \( \theta_i \) representing the weight of topic \( k \) for the \( t \)-th utterance. Here the topic weights of an utterance are not restricted to be non-negative. We assume that \( \theta_i \) depends on \( \theta_{t-1} \), the topic weights of the preceding utterance based on the following relationship,

\[
\theta_i = F_i(\theta_{t-1}) + \epsilon_i = M_i\theta_{t-1} + \epsilon_i.
\]

(1)
Fig. 1: The utterance-level latent topic transition model. $\theta_t$ is a hidden $K$-dimensional topic weight vector of the $t$-th utterance in a spoken document, and $K$ is the number of topics. The observation $x_t$ is a $V$-dimensional vector representing the normalized term frequencies of the $t$-th utterance, and $V$ is the lexicon size. $\theta_{t-1}$ evolves to $\theta_t$ based on a topic transition matrix $M_t$, while the observations $x_t$ is generated from the hidden vector $\theta_t$ through a matrix $C$. $M_t$ is a $K \times K$ matrix, and $C$ is a $V \times K$ matrix.

$F_t(.)$ is the latent topic transition function modeling the relation between $\theta_{t-1}$ and $\theta_t$. Although $F_t(.)$ can be of any form, here we assume the relation is linear and modeled by a $K \times K$ matrix $M_t$, $\epsilon_t$ is a $K$-dimensional prediction error vector absorbing the difference between $\theta_t$ and $M_t\theta_{t-1}$. The transition matrix $M_t$ can be defined based on some heuristic assumptions or trained with a document set, as will be clear in Section 3.

The relation between the latent topic weights and the normalized term frequencies for an utterance is formulated as

$$x_t = C\theta_t + \epsilon^\theta_t,$$

(2)

where $x_t$ is a vector with dimension $V$ which is the size of the lexicon, representing the normalized term frequencies for the $t$-th utterance. The $i$-th dimension of the vector $x_t$ is the normalized term frequency of the $i$-th term $w_i$:

$$x_t[i] = \frac{n(s_i,w_i)}{\sum_{w_j \in x_t} n(s_i,w_j)}.$$  

(3)

$s_i$ is the transcription of the $i$-th utterance in the spoken document (either 1-best or in lattice form), and $n(s_i,w_i)$ is the frequency of term $w_i$ in $s_i$. When $s_i$ is in the lattice form, the occurrences of the term should be weighted by recognition scores such as the confidence measure. The matrix $C$ in (2) is a $V \times K$ matrix that models the term-topic co-occurrence relationships. $\epsilon^\theta_t$ is a $V$-dimensional error vector absorbing the difference between $C\theta_t$ and $x_t$. Based on (2),

$$x_t[i] = \sum_{k=1}^K C[i,k]\theta_t[k] + \epsilon^\theta_t[i],$$

(4)

where $C[i,k]$ is the $[i,k]$ element of the matrix $C$, and $\epsilon^\theta_t[i]$ is the $i$-th element of $\epsilon^\theta_t$. Based on (4), the value of $x_t[i]$, the normalized term frequency for term $w_i$, is the multiplication of $C[i,k]$ and $\theta_t[k]$ summed over the $K$ latent topics with error $\epsilon^\theta_t[i]$. $C[i,k]$ can be understood as the normalized term frequency of term $w_i$ in an utterance given a unit weight of latent topic $k$.

The matrix $C$ above can be obtained in different ways. In the experiments reported below, $C$ is obtained based on utterance-level Probability Latent Semantic Analysis (PLSA) [6], which uses a set of latent topic variables $\{T_k, k = 1, 2, ..., K\}$ to characterize the “term-topic” co-occurrence relationships. Given a training utterance set, PLSA training yields $\{P(w_i|T_k), i = 1, 2, ..., V, k = 1, 2, ..., K\}$, the probability of observing the term $w_i$ in an utterance given the topic $T_k$, and $\{P(T_k|x), k = 1, 2, ..., K\}$, the mixture weight of topic $T_k$ for all the utterance transcriptions $x$ in the training set. This is accomplished by the EM algorithm for maximizing a likelihood function. Here $C[i,k]$ used in (2) and (4) is simply set to the probabilities $P(w_i|T_k)$ obtained from PLSA.

The problem now is to find the hidden sequence of topic weight vectors $\{\theta_1, \theta_2, ..., \theta_T\}$ for a spoken document of $T$ utterances given the observation sequence of normalized term frequency vectors $\{x_1, x_2, ..., x_T\}$, whereas the latent topic weights $\theta_t$ for each utterance and the error vectors $\epsilon^\theta_t$ and $\epsilon^\theta_t$ are all hidden. To make the problem tractable, we assume $\epsilon^\theta_t$ and $\epsilon^\theta_t$ are both sample vectors generated from zero-mean normal distributions respectively with dimensions $K$ and $V$, and covariance matrices $\Sigma_\theta$ and $\Sigma_\epsilon$. In other words, although we never know the true latent topic weights generating the utterances, based on the assumption of the probability distributions for the error vectors, we can estimate the most possible topic weight sequence $\{\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_T\}$ given the observations $\{x_1, x_2, ..., x_T\}$.

With (1), (2) and the assumptions about the distributions generating $\epsilon^\theta_t$ and $\epsilon^\theta_t$, we have

$$P(\theta_t|\theta_{t-1}) \propto \exp\left\{-\frac{1}{2} [\theta_t - M_t\theta_{t-1}]^T \Sigma_\theta^{-1} [\theta_t - M_t\theta_{t-1}] \right\} (2\pi)^{-K/2} |\Sigma_\theta|^{-1/2},$$

(5)

$$P(x_t|\theta_t) \propto \exp\left\{-\frac{1}{2} [x_t - C\theta_t]^T \Sigma_\epsilon^{-1} [x_t - C\theta_t] \right\} (2\pi)^{-V/2} |\Sigma_\epsilon|^{-1/2},$$

(6)

and the joint probability for a sequence of normalized term frequencies $\{x_1, x_2, ..., x_t\}$ and a sequence of topic weight vectors $\{\theta_1, \theta_2, ..., \theta_t\}$,

$$P(\{x_1, x_2, ..., x_T\}, \{\theta_1, \theta_2, ..., \theta_T\}) = \prod_{n=1}^t P(\theta_n|\theta_{n-1}) \prod_{n=1}^t P(x_n|\theta_n).$$

(7)

When the topic transition matrix $M_t$ and the term-topic co-occurrence relationship matrix $C$ are given, it is possible to find the topic weight vector $\hat{\theta}_t$ for the $t$-th utterance given the observations $\{x_1, x_2, ..., x_t\}$ such that

$$\hat{\theta}_t = \arg\max_{\theta_t} P(\theta_t|x_1, x_2, ..., x_t),$$

(8)

where the value of $P(\theta_t|x_1, x_2, ..., x_t)$ is obtainable based on (7). $\hat{\theta}_t$ is the best estimate for the topic weight vector $\theta_t$ of the $t$-th utterance with maximum posterior probability given observations $x_1$ to $x_t$. Equation (8) here is actually solved by the algorithm of Kalman filtering [7].

3. DIFFERENT TOPIC TRANSITION MATRIX

Here we assume two forms of the topic transition matrix $M_t$ as presented below.

3.1. Identity Transition Matrix

Since language is intrinsically cohesive and coherent, the neighboring sentences are usually about very similar topics; hence it is reasonable to assume that the topic weight vectors of adjacent sentences...
are nearly the same. Therefore, we may set the topic transition matrix \( M_t \) equal to the identity matrix for all \( t \) (\( M_t = I \)), and the variations in the topic weight vectors of adjacent sentences are thus all absorbed by the error vectors \( e_t \) in (1). The topic weight vector for the \( t \)-th utterance obtained by (8) with the assumption that the transition matrix is identity is denoted as \( \hat{\theta}^t_0 \), which can be used as the initial topic weight vector for estimation in the next subsection.

3.2. Estimated Transition Matrices

If we have a training spoken document set \( D \) and initial topic weight vectors of all utterances in this set \( D \), we can estimate the transition matrices iteratively using these data. We first cluster all the utterances in the document set \( D \) in an unsupervised way. In the experiments below, we used k-means over the PLSA topic weight vectors for clustering. We assume that the utterances in each cluster thus obtained share the same transition matrix, that is, if there are \( R \) clusters for the document set \( D \), we should have \( R \) different transition matrices, one for each cluster. At iteration \( i \), a set of transition matrices \( \{ M'_1, \ldots, M'_i, \ldots, M'_R \} \) is estimated such that \( M'_r \) for cluster \( r \) minimizes the prediction error over all utterances belonging to the cluster \( r \) in \( D \) when (1) is applied:

\[
M'_r = \arg \min_M \sum_{d \in D} \sum_{s_{t-1} \in d} C(s_{t-1}) = r \left( \hat{\theta}^{t-1}_r - \hat{\theta}^{t-1}_0 \right)^T \left( \hat{\theta}^{t-1}_0 - M'_r \hat{\theta}^{t-1}_0 \right),
\]

where \( C(s_{t-1}) \) represents the cluster ID for the transcription of the \( t - 1 \)-th utterance, or the cluster which \( s_{t-1} \) belongs to, \( \hat{\theta}^{t-1}_0 \) is the topic weight vector of \( s_{t-1} \) obtained at the iteration \( t - 1 \), and the initial topic weight vector \( \hat{\theta}^{t-1}_0 \) is already obtained in Section 3.1. The second summation in (9) is over all utterances in the document \( d \) belonging to the cluster \( r \), and the first summation is over all documents \( d \) in the set \( D \). After obtaining the \( R \) transition matrices, the topic transition in (1) is modified into

\[
\theta_t = M'_r \theta_{t-1} + e_t^r.
\]

That is, the transition matrix from \( \theta_{t-1} \) to \( \theta_t \) is \( M'_r \), and \( C(s_{t-1}) \) is the cluster \( s_{t-1} \) belongs to. Then Kalman filtering is used to estimate the topic weight vectors for all utterances as in (8), which can then be used as the topic weight vectors \( \hat{\theta}^t_r \) for estimating the transition matrices in the next iteration. After \( N \) training iterations, the topic weight vectors \( \hat{\theta}^N_r \) are further used for summarization in the next section.

4. SPEECH SUMMARIZATION WITH SENTENCE-LEVEL LATENT TOPICS

In the initial experiments on speech summarization, we used the Maximum Marginal Relevance (MMR) score [8]. This approach selects in each iteration one utterance from the document to be added to the summary, which is the utterance with the highest similarity to the whole document, while adding minimum redundancy to the summary. This is achieved at each iteration by evaluating a MMR score for each utterance \( s_n \) which has not been added to the summary, and then the utterance with the highest MMR score is selected,

\[
\text{MMR}(s_n) = \lambda S(s_n, d) - (1 - \lambda) S(s_n, d_{sum}),
\]

where \( S(s_n, d) \) is the similarity measure between \( s_n \) and the whole document \( d \), \( d_{sum} \) is the summary obtained in the current iteration, and

\[
S(s_n, d) = \frac{1}{|d|} \sum_{s_b \in d} SIM(s_n, s_b),
\]

where \( SIM(s_n, s_b) \) is the similarity measure between \( s_n \) and \( s_b \), and \( |d| \) is the number of utterances in \( d \). Therefore, the first term on the right hand side of (11) is to be maximized, while the second term is to be minimized, and the parameters \( \lambda \) is to properly weight these two goals.

In the experiments to be reported below, three different approaches were used to estimate the similarity \( SIM(s_n, s_b) \) in (12) all based on the cosine similarities between the vector representations \( v_a \) and \( v_b \) for \( s_n \) and \( s_b \). These three approaches are respectively referred to as word-based, PLSA-based, and transition-based here. For word-based similarity \( SIM_{\text{word}}(s_n, s_b) \), each component of \( v_t \) corresponds to a word in the lexicon, whose value is the term frequency weighted by the latent topic entropy for the term [9]. For PLSA-based similarity \( SIM_{\text{plsa}}(s_n, s_b) \), the dimension of \( v_n \) is the number of latent topics \( K \), and the value of each component is simply \( P(T_k|s) \) from PLSA. For transition-based similarity \( SIM_{\text{trans}}(s_n, s_b) \), \( v_n \) can be either \( \hat{\theta}^{N-1}_r \) obtained above in Section 3.1 or \( \hat{\theta}^{N}_r \) obtained in Section 3.2 with different cluster numbers \( R \). Different similarity measures can be further integrated.

5. EXPERIMENTS

5.1. Experimental Setup

The corpus used in this research is the lectures for a course offered at National Taiwan University. The lectures were given in the host language of Mandarin Chinese but with many technical terms uttered in the guest language of English. The lectures covered a total of 17 chapters, with a total length of 45.2 hours. The corpus was segmented into 193 documents based on the slides used. The average document length was about 17.5 minutes. One-best ASR transcriptions were used for testing. For ASR, the acoustic models were trained on the ASTMIC corpus for Mandarin and on the Simica Taiwan English corpus for English (both included hundreds of speakers), and then adapted using a 25-minute bilingual corpus from the target speaker (the course instructor). The language model was trained with a set of corpus in Chinese, and then adapted with two other courses offered by the same instructor and the course slides.

The accuracies for the ASR transcriptions were 78.15% for Mandarin characters, 53.44% for English words, and 76.26% overall. The reference summaries of 40 documents were provided by graduate students who had taken this course. Only these 40 documents were used for testing. The utterances in all documents were used in PLSA and transition matrix training. The transition matrix training was performed with 5 iterations (\( N = 5 \)). The ROUGE-1, 2, 3 and L F-measures with summarization ratios 5%, 10%, and 15% obtained from the package ROUGE [10] were used to evaluate the summarization results. In the experiments, both \( \Sigma_a \) and \( \Sigma_d \) for \( e_t^r \) in (1) and (2) were set to \( \sigma^2 I \) with \( \sigma \) set to 0.01. The number of latent topics \( K \) was set to 16. \( \lambda \) for MMR in (11) was 0.9.

5.2. Experimental Results

Table 1 lists the ROUGE-1, 2, 3 and L results with 5%, 10% and 15% summarization ratios using different sentence similarity measures \( SIM(s_n, s_b) \) for the MMR scores in (12). Column (A) shows the results for word-based similarity, or \( SIM_{\text{word}}(s_n, s_b) \) was used.
for $SIM(s_a, s_b)$ in (12). Column (B) is the results when PLSA-based similarity was used in addition, or $S_{word}(s_a, s_b)S_{plsa}(s_a, s_b)$ was used for $SIM(s_a, s_b)$ in (12). Column (C) is the results further integrating the information considering topic transition, or $S_{word}(s_a, s_b)S_{plsa}(s_a, s_b)S_{trans}(s_a, s_b)$ was used for $SIM(s_a, s_b)$ in (12). Columns (C1), (C2), and (C3) under column (C) are the results using different topic transition matrices. Paired t-test with significance level at 0.05 was used to test the significance of the improvements obtained. The superscripts $\alpha$, $\beta$, $\gamma$ and $\delta$ respectively represent significantly better than the results in columns (A), (B), (C1) and (C2) on the same row. The ROUGE-1 of the “random”, “head” and “longest” baselines with 5% summarization ratio are 0.261, 0.244 and 0.228 respectively.

<table>
<thead>
<tr>
<th>(A) word +PLSA+Topic Transition</th>
<th>(B) word +PLSA</th>
<th>(C) Estimated</th>
<th>(C1) Iden. Matrix</th>
<th>(C2) 1 cluster</th>
<th>(C3) 2 clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROUGE-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>0.292</td>
<td>0.289</td>
<td>0.296</td>
<td>0.295</td>
<td>0.298</td>
</tr>
<tr>
<td>10%</td>
<td>0.384</td>
<td>0.395</td>
<td>0.400</td>
<td>0.404</td>
<td>0.406$^\alpha$</td>
</tr>
<tr>
<td>15%</td>
<td>0.425</td>
<td>0.443$^\alpha$</td>
<td>0.440</td>
<td>0.446$^\alpha$</td>
<td>0.458$^{**}$</td>
</tr>
<tr>
<td>ROUGE-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>0.121</td>
<td>0.115</td>
<td>0.129</td>
<td>0.123</td>
<td>0.131</td>
</tr>
<tr>
<td>10%</td>
<td>0.162</td>
<td>0.173</td>
<td>0.179</td>
<td>0.180</td>
<td>0.189$^\alpha$</td>
</tr>
<tr>
<td>15%</td>
<td>0.197</td>
<td>0.211</td>
<td>0.209</td>
<td>0.217</td>
<td>0.240$^{**}$</td>
</tr>
<tr>
<td>ROUGE-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>0.085</td>
<td>0.078</td>
<td>0.092</td>
<td>0.084</td>
<td>0.097</td>
</tr>
<tr>
<td>10%</td>
<td>0.112</td>
<td>0.125</td>
<td>0.131</td>
<td>0.131</td>
<td>0.140$^\alpha$</td>
</tr>
<tr>
<td>15%</td>
<td>0.146</td>
<td>0.156</td>
<td>0.157</td>
<td>0.163</td>
<td>0.187$^{**}$</td>
</tr>
<tr>
<td>ROUGE-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>0.285</td>
<td>0.281</td>
<td>0.291</td>
<td>0.288</td>
<td>0.295</td>
</tr>
<tr>
<td>10%</td>
<td>0.377</td>
<td>0.388</td>
<td>0.393</td>
<td>0.397</td>
<td>0.399$^\alpha$</td>
</tr>
<tr>
<td>15%</td>
<td>0.417</td>
<td>0.435$^\alpha$</td>
<td>0.432</td>
<td>0.438$^\alpha$</td>
<td>0.451$^{**}$</td>
</tr>
</tbody>
</table>

Table 1: The ROUGE-1, 2, 3 and L results with 5%, 10% and 15% summarization ratios using different utterance similarity measures $SIM(s_a, s_b)$ in (12). Paired t-test with significance level at 0.05 was used to test the significance. The superscripts $\alpha$, $\beta$, $\gamma$ and $\delta$ respectively represent significantly better than the results in columns (A), (B), (C1) and (C2) on the same row. The ROUGE-1 of the “random”, “head” and “longest” baselines with 5% summarization ratio are 0.261, 0.244 and 0.228 respectively.

Next consider columns (C1), (C2) and (C3) using the latent topic transition model. Comparing column (C1) using the identity matrix with columns (A) and (B), we note that latent topic transition model was always better in all the evaluation metrics for 5% and 10% summarization ratios, even if only the identity matrix was used. Next, columns (C2) and (C3) are the results of using estimated transition matrices. Only the results of using 1 and 2 clusters, or using 1 and 2 latent topic transition matrices are shown here for space limitation. We notice that there is no significance difference between using a single estimated matrix (column (C2)) and the identity matrix (column (C1)), probably because both of them used only a single topic transition matrix, and when only one transition matrix can be estimated, this transition matrix estimated from the training data set turned out to be somewhat similar to an identity matrix, which is reasonable as mentioned above. Moreover, the single estimated matrix (column (C2)) outperformed PLSA (column (B)) in all cases although the improvements were not significant. When we further extended the cluster number to 2, we see with 2 transition matrices to model the topic transitions in column (C3), the results outperformed the identity matrix (column (C1)) and the single matrix (column (C2)) in all cases. Also it can be found that column (C3) was significantly better than columns (A), (B), (C1) and (C2) in many cases (indicated by superscriptions $\alpha$, $\beta$, $\gamma$ and $\delta$). For example, (C3) was significantly better than column (C2) for ROUGE-2 and ROUGE-3 with 15% summarization ratio. This verified that the proposed latent topic transition model offered more accurate latent topic weights behind each utterance.

6. CONCLUSION

In this paper, we propose to model the latent topic transition between adjacent utterances in a spoken document via a topic transition matrix. We show that this latent topic transition modeling offered improvements in speech summarization. Also, a set of transition matrices estimated for clustered utterances in a training document set turned out to yield very good results.

7. REFERENCES