News Topic Specific Language Model Based on Spectral Clustering and Web Crawling *

Shinya Takahashi, Tsuyoshi Morimoto and Naoyuki Tsuruta †

Abstract— **Recently, several adaptation methods for speech recognition language models using a large amount of documents in the World Wide Web (WWW) have been proposed. These methods collect topic-related documents from WWW site and adapt the general language model to a topic specific language model. The topic specific language model can give high performance if the large amount of the appropriate topic-related documents can be collected. However, if the sufficient amount of the appropriate documents cannot be collected, the adaptation does not work well. In this paper we propose a language model adaptation method based on spectral clustering in order to select the appropriate topic-related documents from Web site. To show the effectiveness of this approach, spectral clustering and speech recognition experiments are demonstrated for broadcast news speech. Experimental results show the proposed method can select the suitable cluster that consists of the topic-related documents and can improve the speech recognition performance.**

Keywords: topic specific language model, spectral clustering, speech recognition, Web mining

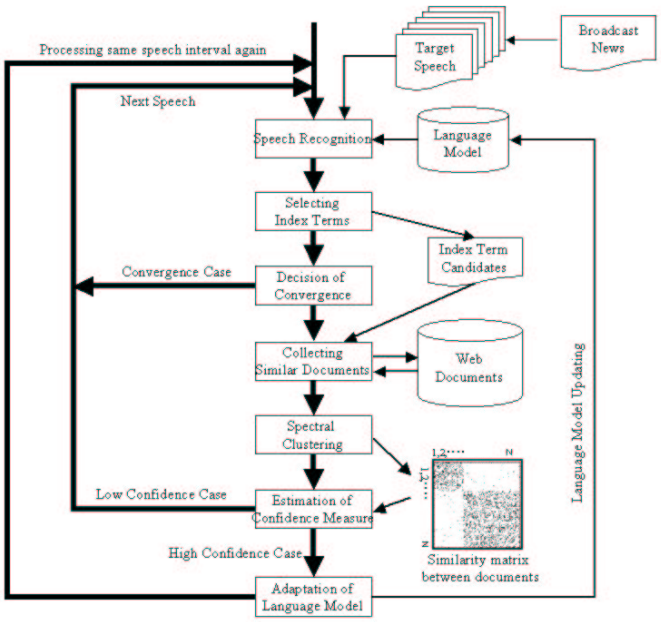

Figure 1: Index term automatic extraction system.

1 Introduction

The World Wide Web (WWW) has been quite wide spread out all over the world and a tremendous number of Web pages has been accessible. These rapid developments of WWW make it possible to utilize excellent and enormous textual data resources for creating topic specific language models.

On the other hand, broadband progress of the Internet makes it possible to provide the large amount of multimedia information on the WWW. Recently, we can watch programs almost same as broadcast TV programs. In addition, recent remarkable advance of personal computers enables users to store these mass data not only from the Internet but also from TV directly. It is need to bind some index terms that represent its contents appropriately to a search any data required by users from these enormous data.

Under these circumstances, we have been developing the broadcast news search system with the language model adaptation using the information on the WWW. The basic idea is that broadcast news have similar Web documents on the Internet news site, so the performance of news speech recognition can be improved with the adapted language model by collecting the similar article via Web crawling [1], [2].

Several methods collecting the topic-related documents automatically have been proposed [3]– [8]. For example, the method proposed in [4] collects similar documents from Web site with previous user utterances as queries to search engine and adapts a general language model for unknown words. Almost of the other methods, such as [7], require to select an appropriate keyword as a query term manually or using some other methods. On the other hand, a unsupervised language model adaptation method based on "query-based sampling" has been proposed [8]. This method selects the keywords from recognition result directly and collects topic-related documents with these keywords from WWW for topic adaptation.

*Manuscript received May 31, 2007.

†Dept. EECS, Fukuoka University, 8-19-1 Nanakuma, Jonan-ku, Fukuoka-shi, 814-0180 Japan (Tel: +81-92-871-6631(ext.6572); Fax: +81-92-871-6031; Email: takahasi@tl.fukuoka-u.ac.jp).

The topic specific language models can give high performance for speech recognition if the large amount of appropriate topic-related documents can be collected. However, if the expected documents cannot be collected, the adaptation does not work well.

In this paper we propose a document filtering method for the language model adaptation using a spectral clustering approach in order to select the appropriate topic-related documents. To show the effectiveness of this approach, speech recognition experiments are demonstrated for the broadcast news speech. Experimental results show the proposed method can give the suitable cluster that consists of the topic-related documents and can improve the speech recognition performance.

2 System Overview

2.1 Automatic extraction of index terms

Figure 1 shows an overview of the index term extraction system we have been developing now. This system processes an input news speech document as follows,

1. A news speech document is recognized using a large vocabulary speech recognizer Julius [9] with a general n-gram prepared in advance.
2. Candidates of keywords are selected from high frequency words in a recognition result.
3. Web documents similar to the recognition result are retrieved using a search engine with the keywords.
4. Text corpus is extracted from the collected web documents using spectral clustering.
5. A topic specific n-gram language model is trained from the extracted corpus.
6. The input news speech document is recognized with the topic specific n-gram.

The above procedure is executed iteratively until the index terms convergences.

2.2 Collecting topic-related documents

As mentioned before, the most critical issue in this approach is the reliability or confidence of the recognition result on the 1st stage. If there are a lot of mis-recognition words, the system may collect a lot of dis-similar documents from Web site. To cope with this issue, we investigate to evaluate the confidence of the recognition result using the collected documents.

The basic idea is that higher confidence results can collect more amounts of similar documents because there would be a lot of news articles similar to the input news speech documents. To the contrary, lower confidence results are expected to collect dis-similar articles, so there would be several dis-similar clusters in the collected documents. So we consider applying cluster analysis to the collected documents.

2.3 Constructing topic-related corpus

In the similar document retrieval stage, we chose the nouns, such as generic nouns, proper nouns, and verbal nouns, from the recognition result as the keywords and retrieved the documents including these keywords using Yahoo! Japan[1] as the search engine. Here, the number of the documents including all keywords might be small, so we perform the retrieval process deleting the lowest frequency word from the keywords iteratively until we obtain sufficient amount of the documents.

After collecting the documents, we calculate the similarity matrix between the documents using cosine similarity with TF weights in a vector space model [10] [2]. Then, spectral clustering, which is explained later, is used for cluster analysis.

2.4 Language model updating

There are two methods for merging a topic specific language model and a general language model. One is a method that concatenates a training corpus and a general corpus [8] simply. The other is a method that merges and assigns a new n-gram constructed from the training corpus to the general n-gram [11]. However, both method require a reliable and a large amount of resources, so it is difficult to apply to a small task such that we deals with in this paper.

So, in order to deal with unknown words, we merged a topic specific dictionary with a general dictionary and we used the topic specific n-gram constructed from the collected documents with the merged dictionary. In the experiments described later, we used the method that merges two language models for comparison.

3 Document Filtering based on Spectral Clustering

3.1 Spectral clustering

Spectral clustering is one of the clustering method that uses eigenvectors of the Laplacian of the symmetric matrix $\mathbf{W} = (w_{ij})$ containing the pairwise similarity between data objects i, j [12]. Spectral clustering has been proposed from graph theory field to solve a minimum cut set problem of non-directional graph. In recent years this method has been applied to many applications, such as

[1] http://search.yahoo.co.jp

[2] The reason why we don't use IDF is that IDF tends to depend on the target domain.

document clustering [13], image recognition [14] and so on.

Given n objects and the similarities between them $W = (w_{ij})$, consider 2-way clustering problem that decide membership indicator as following,

$$q_i = \begin{cases} 1 & \text{if } i \in A, \\ -1 & \text{if } i \in B. \end{cases}$$

Here, if nodes i and j are connected, $w_{ij} = 1$, otherwise $w_{ij} = 0$.

This problem is defined as the minimization problem of the following equation,

$$J(\mathbf{q}) = \frac{1}{2}\mathbf{q}^T(D - W)\mathbf{q},$$

where D is a diagonal matrix with each diagonal element being the sum of the corresponding row ($d_i = \sum_j w_{ij}$).

Using a Lagrangian multiplier under the constraint $\mathbf{q}^T D\mathbf{q} = 1$, and relaxing the restriction of q_i from discrete values to continuous values in $(-1, 1)$, the minimization of $J(\mathbf{q})$ becomes the following eigenvalue problem

$$(D - W)\mathbf{q} = \lambda D\mathbf{q}. \quad (1)$$

Because elements of $\mathbf{q}$ are already relaxed to continuous value, each q_i takes continuous value that satisfies $-1 \leq q_i \leq 1$. Therefore, each node is classified into two classes that satisfy $\{i|q_i < 0\}$ and $\{j|q_j \geq 0\}$. Here, the eigen vector for the minimum eigen value is optimal solution, because we are going to minimize the cut size J. The first eigen value, however, is zero, and its eigen vector is a direct current component so that meaningful solution is the second eigen vector or later than second.

In this paper, we use the following eigenvalue equation used in [15], substituting $\mathbf{q} = D^{-1/2}\mathbf{z}$ into Eq.(1),

$$D^{-1/2}WD^{-1/2}\mathbf{z} = \lambda'\mathbf{z}, \quad \lambda' = 1 - \lambda. \quad (2)$$

It should be noted that here the solution is the eigen vectors corresponding to the second or later largest eigen value because $\lambda' \leq 1$ for Eq.(2).

The solution to solve the graph partition problem described above can be applied to a general clustering problem by replacing weights between each node into similarities between each data. Basically, a eigen vector gives an optimal partition, so $\log_2 k$ eigen vectors give 2^k cluster boundaries.

3.2 Reordering the similarity matrix

By sorting each element of the 2nd eigen vectors by its value and arranging the target data in the corresponding element order, more similar data can be replaced closer with the topographic relations. In addition, a segmentation with appropriate boundaries enables an approximately clustering. In [15], the cluster boundaries were obtained by detecting edge from a connectivity matrix emphasizing boundaries of the similarity matrix. In our case, however, there might be not clearly independent topic in the collected documents so that we calculate cluster boundaries from the eigen vector directly. So, in this paper, based on the principle of graph partition, we decide the candidates of the cluster boundaries as the points that have low d_i shown in Section 3.1. The point with low d_i means that a connective weight of the point between the graphs is low when a target graph is partitioned into two graphs. Selecting the boundary candidate points in the order of lower d_i, we can obtain cluster regions sandwiched between them. Figure 2 shows an example of spectral clustering for a cosine similarity matrix among documents including three topics. Figure 2(c) was obtained by sorting the elements of the second eigen vector. The reason of mosaic like result is that it is insufficient to use the second eigen vector only for sorting the matrix perfectly.

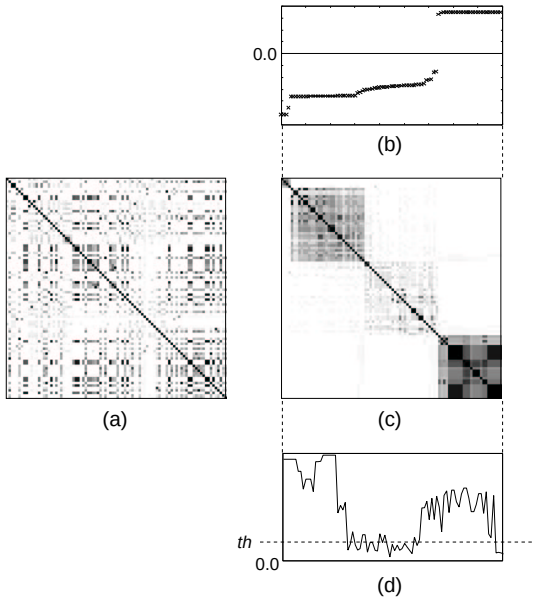

Figure 2: Example of spectral clustering. (a) original cosine similarity matrix among documents including two topics and noises. (b) elements of the second eigen vector. (c) result of reordered matrix. (d) a graph of the difference of adjacent d_i, which is diagonal elements of D.

3.3 Document Filtering Algorithm

The purpose of the proposed approach is 1) to select a cluster that consists of a certain amount of documents similar to each other and 2) to filter out unsuitable documents caused by mis-recognitions. It can be supposed that the document collected with low confidence recogni-

Table 1: News topics used in experiments.

Topic		utt. time	# of utt.	# of words	# of Index term
1	Threat of rainstorm in north Japan	85.8s	7	269	41
2	Washington subway derailment	56.1s	5	186	36
3	Resignation of Polish archbishop *	59.8s	6	174	35
4	Thalidomide as cancer drug **	86.0s	6	275	53

* with background noise by crowd, ** with a press conference interview

tion results tends to consist of some different words from the document collected with high confidence recognition results in same topic.

The complete algorithm is as follows:

1. collect the documents using the recognition results for utterances in each topic,
2. calculate the cosine similarity matrix containing the pairwise similarity between the all documents that consist of recognized sentences and the collected documents,
3. arrange the similarity matrix via spectral clustering,
4. calculate the boundary points using diagonal elements of D in the reordered matrix,
5. extract the target region that is the widest range and includes the most number of utterances from each region sandwiched between the boundary points.

4 Experiments

4.1 Speech data and speech recognition tools

We conducted speech recognition experiments for each utterance separated from news speech by using zero-crossing rate and power of the speech signal. We used 4 topic speech data from NHK TV news program broadcast at 1pm on January 8th, 2007 such as shown in Table 1. Web documents we used in the experiment were collected in a week after the news broadcast.

A bigram language model trained from news paper documents with 20,000 words, which is included in Julius dictation kit Ver.3.1, was used as a general language model to obtain the first recognition results. A news topic specific language model, which is a bi-gram model, was created from the topic-related documents after spectral clustering using CMU-Cambridge SLM Tool kit [16]. To deal with unknown words, we merged the topic specific dictionary with the general dictionary. For an acoustic model, a gender independent HMM was used. We used word correct rate for each sentence, recall and precision of the target noun words for evaluation.

Table 2: Selected clusters.

topic	all documents	clusters	selected documents
1	716	5	272
2	383	2	342
3	503	10	344
4	528	5	405

4.2 Experiments for each topic

4.2.1 Result of spectral clustering for each topic

At first we conducted spectral clustering for each topic independently. Figure 3 shows results of spectral clustering for 4 topics. Arrows in the bottom lines means the location of the document that consists of the recognized sentences and each number in parenthesis shows the number of utterance. Two-way arrows indicates the range of a selected cluster. We executed clustering to maximize the number of cluster and selected a cluster that includes more than a half of the target utterances. Here, the maximum number of cluster is less than 10. Figure 3 shows that the utterance with the lowest score of recall or precision is arranged near the edge of the reordered axis. As a result, appropriate clusters can be extracted without the low score utterances for topic 2, 3, and 4. Table 2 shows the number of all documents, the number of clusters, and the number of the documents in the selected cluster.

4.2.2 Performance of topic specific language model

Next, we conducted speech recognition experiments using the topic specific language model. The results of speech recognition for each topic are shown in Table3. The results with general language model and the results with the language model constructed from all collected documents are shown for comparison. As shown in Table3, the recognition performance for all topics are improved significantly compared with that of general language model. Compared with the results with all documents collected from WWW, the index terms for three topics could be extracted more accurately by using the selected cluster.

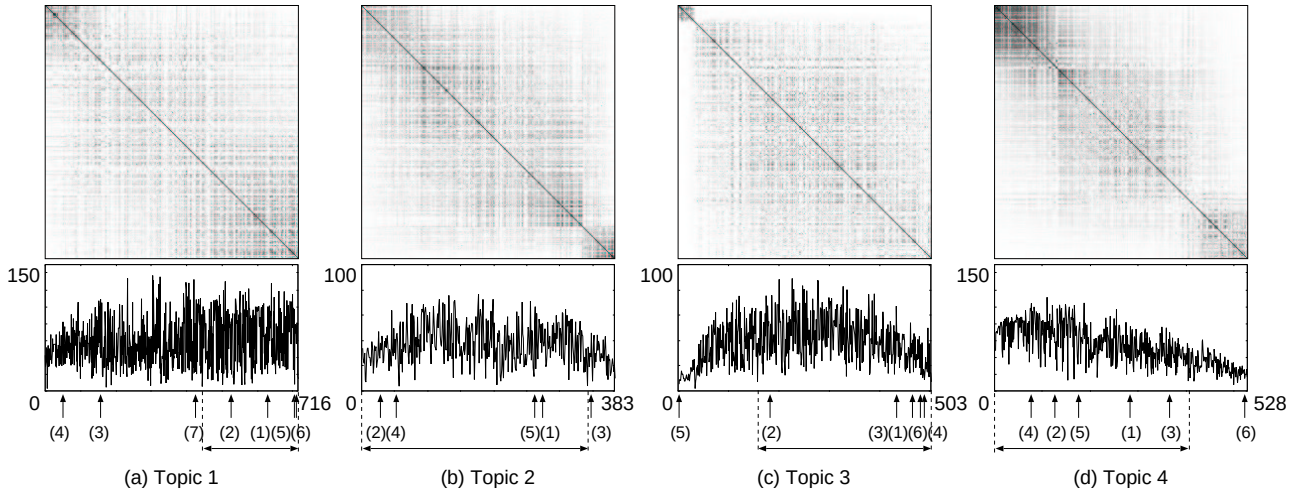

Figure 3: Results of spectral clustering for each topic.

Table 3: Results of speech recognition with news topic specific LM for each topic.

Topic	general LM			Topic specified LM					
				with selected clusters			with all documents		
	WCR	P	R	WCR	P	R	WCR	P	R
1	62.5	44.4	55.8	84.4	81.4	81.4	78.9	69.4	79.1
2	68.8	58.5	66.7	71.5	71.8	77.8	74.2	70.0	77.8
3	72.4	67.5	77.1	74.1	76.3	82.9	73.0	71.8	80.0
4	64.7	51.6	60.4	71.1	65.5	71.7	70.7	65.5	71.7

WCR: Word Correct Rate, P: Precision, R: Recall

4.2.3 Merging to general language model

The results against the merge ratio are shown in Figure 4. Here, "0.0" denotes the result given by the general language model (bi-gram) only and "1.0" denotes that given by the new language model constructed from the selected cluster. Circles, squares and triangles in this graph indicate the word correct rate, the precision, and the recall rate, respectively.

As can be seen from the graph, the result of the topic specific language model without merging to the general language model is better than that of merged language models. This experimental result shows that merging to the general language model is not effective for the target topic in this paper. It is supposed that the reason is that two language models to be merged are so different, as pointed out in [11].

5 Conclusions

In this paper we proposed the document filtering method for the language model adaptation using a spectral clustering approach in order to select the appropriate topic-related documents. From the experimental results of speech recognition for 4 topic broadcast news speech, we confirmed that the proposed method can extract the suitable cluster that consists of the topic-related documents and filter out the unsuitable documents. We showed that the speech recognition performance are improved for all topics using the topic specific language model constructed from the largest cluster that consists of similar documents including the document of the recognized sentences.

References

[1] Takai, D., Morimoto, T., Takahashi, S., " Extraction of index terms for retrieving multimedia news documents from World Wide Web (in Japanese)," *Proc. of the 56th JCEEE Kyushu*, Kumamoto, Japan, 8/03

[2] Takhashi, S., Morimoto, T., Irie, Y., "Adaptation of language model with iterative web crawling for speech recognition of broadcast news (in Japanese)," *Proc. of FIT2006*, Fukuoka Japan, pp. 381–384 9/06

[3] Zhu, X., Rosenfield, R., "Improving trigram language modeling with the world wide web," *Proc. of ICASSP'01*, Salt Lake City, USA, 5/01

[4] Berger, A., Miller, R. "Just-in-time language modeling", *Proc. of ICASSP'98*, Seattle, USA, pp.705–708 12/98

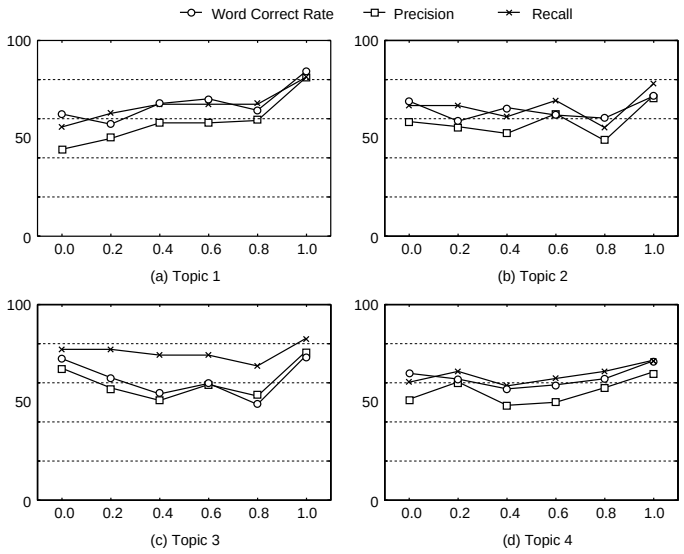

Figure 4: Results with merged language models.

[5] Bulyko, I., Ostendorf, M., Stolcke, A., "Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures", *Proc. of HLT-ACL*, Edmonton Canada, pp.7–9 5/03.

[6] Nishimura, R., *et.al.*,"Automatic N-gram language model creation from web resources", *Proc. of EUROSPEECH-2001*, Aolbarg, Denmark, pp. 2127-2130 9/01

[7] Sethy, A., Georgiou, P.G., Narayanan, S., "Building topic specific language models from webdata using competitive models," *Proc. of INTERSPEECH'06*, Lisboa, Portugal, pp. 1293–1296 9/05

[8] Suzuki, M., Kajiura, Y., Ito, A., Makino, S., "Unsupervised language model adaptation based on automatic text collection from WWW," *Proc. of INTERSPEECH'06*, Pittsburgh, USA, pp.2202–2205 9/06

[9] http://julius.sourceforge.jp/

[10] Salton, G., *et. al.*, "A vector space model for automatic indexing," *Communications of the ACM*, V18, N11, pp.613-620, 1975. Reprinted in *Readings in Information Retrieval*, Jones, K.S. and Willett, P. (Eds.), Morgan Kaufmann Publishers, pp.273-280 1997.

[11] Nagatomo, K., *et.al.*, "Complemental Back-off Algorithm for Merging Language Models (in Japanese)," *IPSJ Journal*, V43, N9, pp.2884-2893 9/02

[12] Pothen, A., Simon, H., Liou, K., "Partitioning sparse matrices with eigenvecotors of graphs," *SIAM J. Matrix Anal.*, V11, N3, pp.430–452, 1990

[13] Dhillon,I.S., "Co-clustering documents and words using bipartite spectral graph partitioning," *ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining*, San Francisco, USA, pp. 269–274 8/01

[14] Tsuruta, N., Aly, S. K. H., Maeda, S., Takahashi, S., Morimoto, T., "Self-organizing Map vs. Spectral Clustering on Visual Feature Extraction for Human Interface," *Proc. of Int. Forum on Strategic Technology (IFOST) 2006*, Ulsan, Korea, pp. 55–58 10/06

[15] Ding, C., He, X., "Linearized Cluster Assignment via Spectral Ordering," *Proc. of ACM Int. Conf. on Machine Learning*, Banff, Canada, pp.30–37 7/04

[16] Clarkson, P.R., Rosenfeld, R., "Statistical Language Modeling Using the CMU-Cambridge Toolkit," *Proc. ESCA Eurospeech*, Rhodes, Greek, pp.2707-2710 9/97