JAPANESE BROADCAST NEWS TRANSCRIPTION AND TOPIC DETECTION

Sadaoki Furui[1], *Koh'ichi Takagi*[1], *Atsushi Iwasaki*[1],
Katsutoshi Ohtsuki[2], *Tatsuo Matsuoka*[3] *and Shoichi Matsunaga*[2]

[1] Tokyo Institute of Technology
2-12-1, Ookayama, Meguro-ku, Tokyo, 152 Japan
[2] NTT Human Interface Labs
1-1, Hikari-no-oka, Yokosuka-shi, Kanagawa, 239 Japan
[3] NTT Multimedia Business Department
2-2-2, Otemachi, Chiyoda-ku, Tokyo, 100 Japan

ABSTRACT

This paper reports recent advances in Japanese broadcast news transcription and automatic topic detection from the transcribed news speech. To cope with the variability of the readings for each word, a new method for incorporating reading probability of each word in the decoding process is proposed. As a realistic solution to the new-word problem, a new method is proposed, in which new words are manually registered and OOV language model is applied to the new word. To detect topic words for news speech, two methods are proposed; one uses a relevance measure between each word in the news and each word in the topic word set, and the other uses a significance measure for each word based on the frequency ratio.

1. INTRODUCTION

We have been conducting research on Japanese broadcast news transcription and have obtained the following results [1]. First, recognition results obtained when the Nikkei (Japanese Economic Journal) newspaper is used for language model training differ greately from those obtained when broadcast news text is used. The broadcast news language model produces much better results than the newspaper language model, although the broadcast news text is much smaller than the newspaper text. Roughly 80% morpheme (word, hereafter) accuracy was achieved for anchor speakers (professional announcers) under the condition that 20k words covering 98% of the most frequent words occurring in the broadcast news text for training were used as the vocabulary, word trigrams were used as language models, triphone HMMs were used as acoustic models, and noiseless clean speech was used for evaluation.

Japanese speech transcription has several Japanese-specific problems, and it is very important to solve these problems for improving the recognition performance. One of the problems is the reading variability of Chinese characters constructing words. Many Japanese words can be read in two or more different ways, and each of the readings has different frequency in use. Often the correct reading can only be determined according to the context. If different readings for each word are treated as different words, the size of the vocabulary becomes huge. Therefore, we have so far made language models of words according to their written forms without considering their readings, and allowed all possible readings with equal probability in the decoding process. If two words using an identical character have different readings and one of them has a high observation probability, the other one will also have high observation probability (likelihood) even if it is hardly used. This sometimes causes recognition errors. This paper addresses this problem.

Technology for automatically detecting words representing topics of utterances is expected to be applied in many different ways, such as making broadcast news databases accessible with keywords. We have been investigating a method of automatically giving topic words to each broadcast news article by using the results of automatic transcription [2]. Instead of simply choosing words from a word set obtained by transcription, we prepared a large-vocabulary topic-word set beforehand, and selected multiple topic words which are most appropriate to the news based on a topic-word detection model. The topic-word detection model indicates a measure of relevance between each word in the news article and each topic word based on the χ^2 distribution. Assuming that the words in

the headline of a newspaper article indicate topics of the article, we trained the relevance measure using words in the newspaper articles and the words in the headlines. This method was evaluated using new articles by comparing the detected topic words with correct topic words given by human subjects. When five topic words were automatically detected using the transcribed broadcast news, 75% of them were correct. This paper investigates a much simpler method in which topic words are directly detected from the transcribed words, and compares the results with those obtained with the above method.

This paper also describes several other investigations we have made on various problems related to broadcast news transcription.

2. STRUCTURE OF THE BROADCAST NEWS TRANSCRIPTION SYSTEM

2.1 Acoustic Models

The feature vector consists of 16 cepstral coefficients, normalized logarithmic power, and their delta features (derivatives). The total number of parameters in each vector is 34. Cepstral coefficients were normalized by the CMS (cepstral mean subtraction) method.

The acoustic models we used were shared-state triphone HMMs designed using tree-based clustering. The total number of states was 2,106, and the number of Gaussian mixture components per state was 4. They were trained using phonetically-balanced sentences and dialogues read by 53 speakers (roughly 20 hours in total). They are completely different from the broadcast news task. All of the speakers were male, thus the HMMs were gender-dependent models. The total number of training utterances was 13,270 and the total length of the training data was approximately 20 hours.

2.2 Language Models

Broadcast-news manuscripts recorded from August 1992 to May 1996, comprising roughly 500k sentences consisting of 22M words, were used for constructing language models. To calculate word n-gram language models, we segmented the broadcast-news manuscripts into words by using a morphological analyzer since Japanese sentences are written without spaces between words. The morphological analyzer is different from that used in our previous work [1]. We had previously used an NTT program, but recently we switched to a public-domain program, "*Chasen*". A word-frequency list was derived for the news manuscripts, and 20k most frequently used words were selected as vocabulary words. The 20k vocabulary covers about 98% of the words in the broadcast-news manuscripts. We calculated bigrams and trigrams, and estimated unseen models using Katz's back-off smoothing method.

2.3 Evaluation Experiments

News speech data, which were broadcast on TV in June 1996, were divided into two parts, an anchor set and an others set, which were separately used for evaluation. The anchor set consists of 100 utterances by five speakers, and the others set consists of 125 utterances by six speakers. All of them are relatively clean speech with no background noise, and they were manually segmented into sentences. This corresponds to the partitioned evaluation (*PE*) with the baseline broadcast condition (*F*0) in the 1996 Hub-4 test. The out-of-vocabulary (OOV) rates are 0.8% and 3.4% for the anchor and others sets, respectively.

Table 1 shows word accuracies and test-set perplexities in the cases of bigrams and trigrams. These results are very similar to the previous results obtained using a different morphological program. This table also shows accuracies using Japanese characters as units. The results for anchors are much better than those for others, which corresponds to the difference of the test-set perplexities.

Table 1: Recognition results using bigrams and trigrams

	Speakers	Test-set perplexity	Word accuracy	Character accuracy
Bigram	Anchor	122.3	76.6 %	83.4 %
	Others	325.5	62.3 %	71.6 %
Trigram	Anchor	61.1	80.7 %	86.4 %
	Others	219.4	65.7 %	73.1 %

1 word ≅ 1.82 characters ≅ 3.96 phonemes

Three subsets of the broadcast-news manuscripts were extracted with some overlaps, and respectively used for calculating bigrams. Table 2 shows OOV rates for the evaluation data under each condition including the case of using whole manuscripts. The OOV rate for anchors increases as the time interval between training and

evaluation increases, and the OOV rate for others are consistently high.

Table 2: OOV rate for the evaluation data under each condition of manuscript used for bigram calculation

News manuscri	Anchor	Others
92/7 – 96/5 (2	0.81 %	3.40 %
95/5 – 96/5 (9.	0.78 %	3.49 %
94/4 – 95/9 (9.	1.06 %	3.08 %
92/7 – 94/9 (9.	1.36 %	3.31 %

Test-set perplexities and word accuracies obtained by using bigrams and trigrams are shown in Fig. 1. Various bigrams calculated as above were used. The results for the bigrams show that, if the time difference between training and evaluation is small, even if the size of manuscript used for calculating the language models is less than half that of a whole manuscript, similar performance to that obtained using a whole manuscript can be achieved. For anchors, both recognition accuracy and test-set perplexity become worse as the time interval between training and evaluation data becomes larger. If trigrams had been calculated by using these reduced manuscripts, the effect of reducing the manuscript would have been more significant than it was for the bigrams.

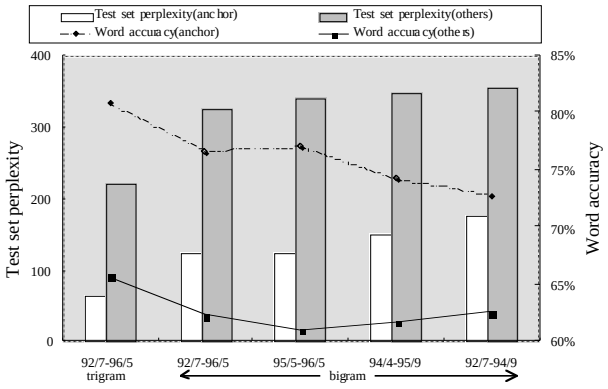

Figure 1: Effects of training manuscript conditions on word accuracy and test-set perplexity (Test set: June 1996).

3. INTRODUCING THE READING PROBABILITY

3.1. Approximation of Language Models Incorporating Reading Probabilities

Table 3 shows three examples of words written with a single Chinese character having various readings. The table includes the probability and meaning for each reading. It should be noted that the meaning sometimes changes greatly according to the reading, and that the probability is highly biased.

Table 3: Reading variation examples.

Written form	Reading	Probability	Meaning
~	eN	0.8234	yen, circle
	tsubura	0.0766	round
	madoka	0.0500	round
	maru	0.0500	circle
	sai	0.9996	on the occasion of
	kiwa	0.0004	edge
{	hoN	0.8138	book, pieces
	boN	0.1000	book, pieces
	moto	0.0862	origin

Suppose k-th word, w_k, in a sentence has reading variations, r_k. Ideally, we should calculate

$$P(w_{k=1}^{n}(r_k)) = \prod_{k=1}^{n} P(w_k(r_k) \mid w_{i=1}^{k-1}(r_i)) \qquad (1)$$

However, as described in the Introduction, these values are very difficult to calculate due to the sparseness problem. Therefore, we introduce the following approximation:

$$P(w_k(r_k) \mid w_{i=1}^{k-1}(r_i)) \approx P(w_k(r_k) \mid w_k) P(w_k \mid w_{i=1}^{k-1}) \qquad (2)$$

The conventional method corresponds to the condition that the first term in the right-hand side of Eq. (2) is fixed at 1.

We calculated the frequency of each reading for each word in the news manuscript using a morphological analysis program, and stored reading probabilities, $P(w_k(r_k)|w_k)$, in the word dictionary. A weighting factor β is given to the probability, and $P(w_k(r_k)|w_k)^{\beta}$ is used as the first right-hand side term of Eq. (2). Since likelihood values are usually calculated in the logarithmic domain, a

term $\beta \log P(w_k(r_k)|w_k)$ is actually added at each transition of word models.

3.2. Evaluation Experiments

Table 4 shows word accuracies obtained when the bigrams calculated using all the news manuscripts were used as language models and the weighting factor b was changed. It is observed that the error rate for anchors is reduced by roughly 5% by setting β at 5.

Table 4: Word accuracy as a function of the weighting factor β for the reading probability $P(w_k(r_k)|w_k)$

	Anchor	Others
0 (Normal)	76.6 %	62.3 %
0.5	76.9 %	62.2 %
1	77.3 %	61.9 %
2	77.5 %	62.5 %
5	77.9 %	63.4 %
7	77.5 %	63.5 %
10	77.0 %	62.7 %

4. NEW WORD REGISTRATION

4.1. New-Word Language Models

Unknown-word problems can not be avoided in broadcast news transcription. However, it is difficult to automatically detect unknown words with the present large-vocabulary continuous-speech recognition technology. Therefore, we decided to employ a realistic solution in which new words were detected by checking previous results of the news transcription and/or by analyzing recent newspaper articles using the dictionary for recognition, and these new words with readings were manually added to the dictionary. We approximate the language model of the new word, *u*, by using the language model of OOV as follows.

$$P(u_k \mid w_{i=1}^{k-1}) \approx P(u_k \mid C_k) P(C_k \mid w_{i=1}^{k-1}) \quad (3)$$

$$\approx \frac{1}{K} P(C_k \mid w_{i=1}^{k-1}) \quad (4)$$

where C is the OOV class, and K is the distinct number of OOV words in the training manuscript.

4.2. Evaluation Experiments

Table 5 shows word accuracies for the case of using bigrams as language models. The results with or without new-word entries are compared. The table includes new-word recognition rates when these entries are given. It is observed that 70 - 80% of new words are correctly recognized and the overall word accuracies are improved by giving new-word entries and approximations of their language models.

Table 5: Effects of new-word language model (bigrams)

	Word accuracy		New-word recognition rate
	New word not given	New word given	
Anchor	76.6%	77.6%	68.8%
Others	62.3%	66.3%	76.7%

5. TOPIC WORD DETECTION

5.1. Method I

We defined a large set of topic words, which are separate from a set of words in the news articles, and a relevance measure between the words in the two sets. The measure was used for detecting a subset of the topic words that are most relevant to each news article (Method I). In our experiment, we used only nouns and verbs, since other parts of speech are considered to be less important in conveying topic information. We extracted words in the headlines of Nikkei newspaper articles published over a five-year period and used them as the topic word set. The following modified χ^2 value was calculated using the Nikkei newspaper database and used as the relevance measure between a vocabulary word w_i and a topic word t_j.

$$\chi^2{}_{ij} = \frac{(f_{ij} - F_{ij}) \mid f_{ij} - F_{ij} \mid}{F_{ij}} \quad (5)$$

$$F_{ij} = \frac{\sum_{l=1}^{M} f_{il} \sum_{k=1}^{N} f_{kj}}{\sum_{k=1}^{N} \sum_{l=1}^{M} f_{kl}} \quad (6)$$

where f_{ij} is the frequency of a word w_i in all the newspaper

articles with a topic word t_j, N is the distinct number of article words (vocabulary words), and M is the distinct number of headline words (topic words). Since only a positive value is meaningful as a measure of relevance, we used the above expression.

In the evaluation using a new broadcast news article, the following value was calculated for each topic word using all nouns and verbs in the article, and the topic words having relatively large values were selected as topic words of the article.

$$\text{score}(t_j) = \sum_{i \in \text{article}} \chi^2_{ij} \qquad (j = 1, 2, ..., M) \tag{7}$$

5.2. Method II

Another method simpler than the Method I was also investigated. In this method, a subset of article words (nouns and verbs), which are considered to represent topics, are detected. Many measures which have been used in information retrieval from text databases were tried in a preliminary experiment, and the following measure was chosen (Method II).

$$\text{score}(w_i) = g_i \cdot \log \frac{G_A}{G_i} \qquad (i = 1, 2, ..., N) \tag{8}$$

where g_i is the frequency of a word w_i in a news article, G_i is the frequency of the word w_i in all the newspaper articles, and G_A is the summation of all G_i's:

$$G_A = \sum_i G_i \tag{9}$$

The same newspaper articles used for the Method I were used for calculating the G_i and G_A values.

5.3 Evaluation Experiments

Performances of topic word detection by Method I and II were compared using the following scores for each broadcast news.

$$\textit{Recall} = (D / T) \bullet 100\ (\%) \tag{10}$$

$$\textit{Precision} = (D / H) \bullet 100\ (\%) \tag{11}$$

where D is the number of correctly detected topic words, T is the total number of correct topic words, and H is the total number of detected topic words.

Twenty-nine broadcast news articles comprising 142 utterances by 15 male speakers (8 anchors and 7 others) were used for evaluation. Each news article has 2 - 14 utterances (5 utterances on the average). Correct topic words were given by three subjects; 10 phrases, which correspond to 24.4 words, were given on the average for each news article by each subject. Two correct topic word sets were constructed for each news; an "AND" set was made from topic words given by all subjects in common (10.4 words on the average for each news article) and an "OR" set was made from topic words given by at least one subject (35.7 words on the average for each news article). Supplementary experiments were also conducted by giving correct texts instead of transcription results as input.

Results for Method I using either the AND set or OR set as the correct topic word set were averaged over the 29 news articles (see Fig. 2). It is observed that, if transcription results are used as input, precision as well as recall is reduced by roughly 10% from that obtained using correct texts as input. It is also shown that there is roughly 20% difference in performance whether the AND set or OR set is used as the correct set.

The precisions obtained by Method I and II under the condition of choosing five topic words for each news article or adjusting the number of topic words so that precision and recall become equal are shown in Table 5, where the OR set was used as the correct set. When five topic words are chosen, there is no significant difference between Method I and II, and precisions of 85% and 75% were obtained for text input and speech input, respectively. When Method I and II are compared under the condition that precision and recall are equal, Method II is better than Method I for both text and speech input. Since there are several other advantages and disadvantages with Method I and II, we need to perform more detailed comparisons under various conditions.

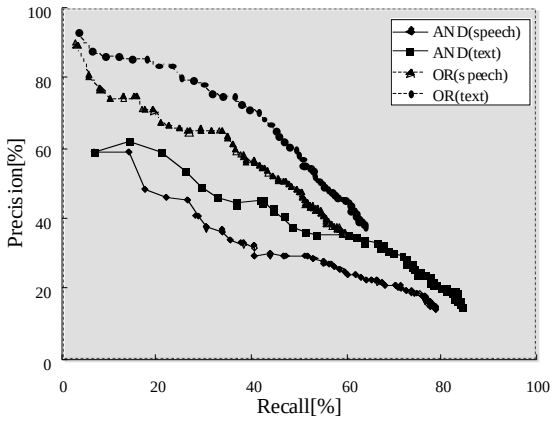

Figure 2: Topic detection from news speech or news text (Method I).

Table 6: The precisions obtained by Methods I and II when 5 topic words are detected or the number of topic words are adjusted so that precision and recall become equal. The OR set is used as the correct set.

	Input	Method I	Method II
Recall = 14.0% (5 topic words)	Text	85.5%	86.5%
	Speech	74.5%	76.1%
Recall = Precision	Text	52.9%	62.0%
	Speech	48.5%	51.2%

6. CONCLUSIONS

This paper reported recent advances in Japanese broadcast news transcription and automatic topic word detection from the transcribed news speech. Word accuracies of 80.7% and 65.7% were obtained for anchors and others, respectively, in the transcription experiments. To cope with the variability of readings for each word, a new method for incorporating the reading probability of each word in the decoding process was proposed, and roughly 5% reduction in error rate was achieved for anchors. Another new method was proposed as a realistic solution to the new-word problem, in which new words are manually registered and OOV language model is used for new words. It was confirmed that 70 - 80% of new words were correctly recognized with this method.

Two methods were proposed to detect topic words for news speech. Method I uses χ^2 measures to select a subset of topic words which are relevant to each news article. Method II uses a measure based on a frequency ratio to detect topic words from the news. When five topic words are detected, both methods achieve a precision of roughly 75%.

Present morphological analysis programs sometimes make errors in giving a reading to each word. Improving the accuracy of giving correct readings in language model training and increasing the size of training manuscript are expected to bring improvement in transcription. Trigrams have so far been used as language models, but it will be worth trying to use higher order statistical language models if we could appropriately cope with the data sparseness problem.

Improvement of the topic word detection performance will be achieved by using multiple hypotheses of speech transcription instead of using only a single hypothesis, which is equivalent to the improvement of transcription accuracy [3]. It will also be helpful to incorporate semantic relationships between words represented by thesauruses and so forth.

ACKNOWLEDGMENTS

The authors would like to thank NHK (Japan Broadcasting Corporation) for providing us with broadcast-news database. The authors are also grateful to Nihon Keizai Shinbun Incorporated for allowing us to use the newspaper text database (Nikkei CD-ROM 90-94) for our research.

REFERENCES

1. T. Matsuoka, Y. Taguchi, K. Ohtsuki, S. Furui and K. Shirai: "Toward automatic transcription of Japanese broadcast news", *Proc. Eurospeech'97*, pp. 915-918 (1997)

2. K. Ohtsuki,T. Matsuoka, S. Matsunaga and S. Furui: "Topic extraction based on continuous speech recognition in broadcast-news speech", *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 527-534 (1997)

3. K. Ohtsuki, T. Matsuoka, S. Matsunaga and S. Furui: "Topic extraction with multiple topic-words in broadcast-news speech", Proc. ICASSP 98 (1998) (to be published)