PERFORMANCE EVALUATION OF WORD PHRASE AND NOUN CATEGORY LANGUAGE MODELS FOR BROADCAST NEWS SPEECH RECOGNITION

Kazuyuki TAKAGI Rei OGURO Kenji HASHIMOTO Kazuhiko OZEKI

The University of Electro-Communications
Chofu, Tokyo, 182-8585 Japan
http://www-oz.cs.uec.ac.jp/

ABSTRACT

This paper reports our work to improve a bigram language model for Japanese TV broadcast news speech recognition. First, frequent word strings were grouped into phrases in order that the phrases were added to the lexicon as new units of recognition. The test set perplexity was improved when frequent function word strings were used as additional recognition units. The speech recognition performance was improved both by grouping function word strings and by grouping compound nouns that were selected by word association ratio. Secondly, in order to alleviate the OOV problem related with nouns, we built and tested a language model that allows switching its noun lexicon according to the domain of the article to be recognized next.

1 INTRODUCTION

The DARPA Hub-4 launched in 1995 is evaluating LVCSR systems to transcribe audio recordings of broadcast news [1]. The evaluation results have been updated annually [2]. The Japanese counterpart of HUB-4 broadcast news materials has been being developed by NHK (Japan Broadcasting Corporation) since 1996 [3]. The broadcast news material consists of audio recordings of NHK TV news programs and manuscript texts. Some of the preliminary recognition results have been reported [4, 5].

Among various challenging issues in automatic transcription of broadcast news, this paper deals with the language modeling. The vocabulary size of broadcast news article is huge and words are from various domains. Also, the topic of articles, thus its vocabulary, changes frequently because of the nature of broadcast news show.

In this paper, two main lines have been pursued to improve a bigram language model: construction of phrases and a category-based model for nouns. The first topic is the concatenation of words into phrases. Grouping frequent word strings into phrases can improve the language model [6, 7, 8]. Based on some linguistic knowledge and intuition, we got started by selecting two types of word strings to be grouped to recognition units: function word strings and compound nouns. Secondly, we attempted to build a language model that allows switching its noun lexicon according to the domain of the article to be recognized next.

2 BROADCAST NEWS DATA

The broadcast news database used in the experiments of this paper was provided by NHK [3], which contains the manuscripts and the audio recordings of a morning news program and an evening news program. The manuscript corpus is a collection of broadcast news articles written by newspersons between August 1992 and August 1996 (40 months, 545K sentences, 93M characters) [1]. Speech corpus consists of speech materials sampled from June 1 to July 14, 1996 (44 days; approx. 45 minutes per day) together with transcriptions. The evaluation set contains 200 sentences sampled from the broadcasts of the period between July 11 to July 14 1996, among which only clean speech portions spoken by news anchors and reporters, with no background noise were used for evaluation.

3 LANGUAGE MODEL

3.1 Baseline Model

The baseline bigram language model (BSL) was trained on the manuscript texts of the first 37 months (478K sentences, 42.9M words, 97.1M characters). Words are not clearly delimited in Japanese text because there is no spacing between words. The manuscript texts were first split into morphemes [2] by morphological analyzer

[1] Manuscripts are proofread and modified by the announcers and the directors before they are on-air. There are usually discrepancies among the newsperson manuscripts, the announcer manuscripts and the transcriptions of the speech actually spoken.

[2] We refer to morphemes as words in LM hereafter.

JUMAN [9], after some preprocessing to filter out titles, headers and annotation marks. According to a word frequency list, 20k most frequently used words were selected as the baseline vocabulary (coverage: 97.75%). Then the bigram language model was generated using CMU-Cambridge SLM [10]. We applied Good-Turing backoff smoothing method (cutoff=1).

3.2 Word Phrase Model

Most Japanese LVCSR systems use morpheme as a recognition unit. However it is questionable whether a morpheme is the best basic unit for stochastic language models. Several reports indicated that grouping frequent word strings into new words may improve the language model performance [6, 7, 8].

3.2.1 Function Word String

Looking at the news manuscript corpus, we noticed that certain types of morpheme strings occur very frequently. Among those are function word strings. For instance, formal and politeness expressions involve fixed patterns of copula, formal verbs and auxiliary verbs. Also, particles that consist of multiple morphemes and function like a single particle are common.

In this paper, words that belong to one of the following seven parts of speech were regarded as function words: formal verb, copula, auxiliary verb, formal noun, demonstrative, particle, suffix. Table 1 shows the number of distinct function word strings whose occurrence frequency in training corpus exceed 100, composed of up to five function words [3].

Each of the selected strings was used as a new recognition unit. Then the function word string model (FWP) was generated whose vocabulary size was determined so that the word coverage ratio is the same as the baseline LM.

Table 1: Selected Function Word Strings

Length	2	3	4	5
#Distinct String	886	1005	469	204

3.2.2 Compound Noun

Secondly, our compound noun model (CN) was designed to refine noun entries, which occupy 61% of the baseline lexicon. In the corpus, while 42% of distinct word 2-tuples are noun 2-tuples; only 15% of them occur more than 5 times. There are noun strings specific to broadcast news speech: for example, some acronyms are always followed by their full notation (e.g. "COD" in Table 2).

However, introducing all the noun 2-tuples as additional lexical entries will in general degrade the quality of the probability estimates of the bigrams. We used word association ratio [11] as a criterion to select frequent noun 2-tuples that appear independently of the domain and of the period of the news article.

$$I(x;y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{1}$$

The noun 2-tuples (x, y) whose association ratio $I(x; y)$ (Equation 1) and occurrence frequency exceed predefined thresholds were selected to give 920 compound noun entries in the lexicon. The bigram language model was generated by the same manner as the function word string model.

Table 2: Selected Compound Nouns (English Translation) and Their Word Association Ratio

$I(x;y)$	x	y
16.5	inside-route train	outside-route train
16.5	Mariana	Trench
16.5	COD	chemical oxygen demand
...	...	...
14.3	yen-selling	dollar-buying
14.1	guilt	arraignment
14.0	assistant	prosecutor

Table 3 shows the test set perplexity and the lexicon size of the baseline bigram model (BSL) and the word phrase bigram models. Compared to the baseline, the function word string model yields significant perplexity reduction while lexicon size was increased 15%. The compound noun model exhibits the same perplexity as the baseline language model.

Table 3: Test Set Perplexity, Lexicon Size of Word-Phrase LMs

LM	BSL	FWP	CN
Test Set PP	89.3	79.3	89.3
Lexicon Size	20k	23k	20k

3.3 Noun Category Model

One of the primary concerns of language modeling is that some words never appear in the training texts. It is natural to assume that the majority of out-of-vocabulary words are nouns especially in broadcast

[3] Function word strings longer than five words are rare.

news. The vocabulary size of broadcast news articles is huge [4] and words are from various domains. Moreover in broadcast news shows, the domain of the article, thus the vocabulary changes frequently. It is also a well-known fact that a set of proper nouns related to a specific event is on-air only within a limited period of time. Our preliminary experiment showed that a recognizer using a dictionary whose noun entries contained the nouns that appeared only in a certain month's articles missed most of the proper nouns in the following month's broadcasts.

In order to alleviate the OOV problem of nouns, we attempted to build a language model, noun category model (NC), that allows switching its noun lexicon according to the domain of the article to be recognized next [5].

Every noun in the training corpus was first replaced by a tag for one of the six noun categories: common, adverbial, temporal, formal, proper, numeral. Then a bigram language model was generated using this preprocessed corpus. It was assumed that conditional word probability within each category is uniform. Then we entered each of 1278 distinct nouns appeared in the evaluation data into the lexicon of the corresponding category. The total lexicon size was 4426.

4 EXPERIMENT

4.1 Recognition System

The system was implemented with HTK [12] as illustrated in Figure 1. A standard acoustic analysis method was employed. Each frame of input speech was represented by a 38 dimensional feature vector that consists of 12 MFCC parameters and their differential coefficients of 1st and 2nd order, together with the differential coefficients of 1st and 2nd order of speech power.

Gender-dependent triphone sets were created using ATR PB Sentences spoken by 201 male and 203 female speakers. The read speech materials were gathered from ATR, ASJ, and JNAS [6] databases. The decoder was run in a single pass performing 50-best recognition, using the bigram language models, with the fixed beam width of 250.

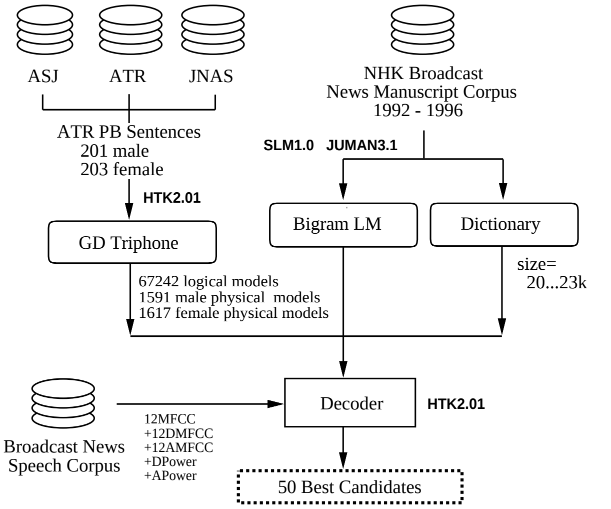

Figure 1: Broadcast News Speech Recognition System

4.2 Experimental Conditions

The evaluation speech data comprising 400 sentences sampled from the broadcasts in July 1996 are divided into four parts: male speaker set (studio clean, others) and female speaker set (studio clean, others), each of which contains 100 sentences. In our experiment, only the studio clean sets were used for evaluation.

The recognition performance was evaluated by word accuracy given in Equation 2, where N is the total number of words, D is the number of deletions, S is the number of substitutions, and I is the number of insertions.

$$Accuracy = \frac{N - D - S - I}{N} \times 100\% \qquad (2)$$

Note that the best grammar scale factor, which post-multiplies the language model likelihoods from the word lattices, for each language model was different. Word accuracy figures in Table 4 were the ones obtained with the best grammar factor for each condition.

5 RESULTS

Both the function word phrase model and the compound noun model improved the performance in their overall word accuracies compared with the baseline model.

The function word string model reduced the substitutions by 10%, because most of the strings of very short function words were correctly recognized as phrases, in speech portions where substitution errors are very likely to be made by the baseline model. On

[4] There are 115.8k distinct words in the training data of our news manuscript corpus.

[5] This is a reasonable design in a practical application, such as automatic subtitling of TV news broadcast, because the articles are arranged beforehand.

[6] Japanese Newspaper Article Sentences, http://www.milab .is.tsukuba.ac.jp/jnas/instruct.html

the other hand, the insertion errors increased by 23%. One of the reasons may be the increased sparsity of training data. Although the test set perplexity was not reduced by compound noun model, the insertion, deletion and substitution errors were reduced by 5.5%, 5.5%, and 10% respectively. Examinations on recognition outputs time-aligned with the transcription showed that compound nouns were recognized as single words correctly.

Finally, the noun category model did not give better word accuracy. Although the lexicon was so generated that it covers nouns that appear in evaluation data, the model lost information about unigram category membership probability of the nouns. Contrary to our expectation, it degraded the overall recognition performance despite the reduction of lexicon size to one fourth that of the baseline.

Table 4: Word Accuracy for NHK Evaluation Set

LM	BSL	FWP	CN	NC
Male	83.9	85.6	84.4	72.0
Female	87.2	87.9	88.3	75.2
Lexicon Size	20k	23k	20k	4.5k

6 CONCLUSION

This paper reported our approaches to improve a bigram language model for Japanese TV broadcast news speech recognition. By using frequent word strings as additional units of recognition, the test set perplexity and the recognizer performance were improved. The compound noun model yielded the best result of 88.3% (50-best) in word accuracy. We also attempted a language model in that nouns are subcategorized allowing to switch its noun lexicon according to the domain of the article to be recognized next. Loss of language constraints by uniform membership within the subcategories may be the reason that the approach did not yield improvements on overall word accuracy. The focus of our future work is on improving the word phrase model by selecting other types of word strings, and by reducing the lexicon size (as some word strings contain shorter ones; some share a substring). As for the noun category model, we are planning to improve it by incorporating within-category unigram.

ACKNOWLEDGMENTS

This work was partially funded by, and the broadcast news database was provided by NHK (Japan Broadcasting Corporation) Science and Technical Research Laboratories.

REFERENCES

[1] A. I. Rudnicky, “HUB4: Business Broadcast News,” Proc. DARPA Speech Recognition Workshop, pp. 8 - 11, Feb. 1996.

[2] D. S. Pallett, J. G. Fiscus, A. Martin, and M. A. Przybocki, “1997 Broadcast News Benchmark Test Results: English and Non-English,” Proc. DARPA Broadcast News Transcription and Understanding Workshop, http://www.nist.gov/speech/proc /darpa98/, Feb. 1998.

[3] A. Ando and E. Miyasaka, “Construction of Japanese News Speech Databases,” Proc. Acoustical Society of Japan Spring Meeting, 2-Q-9, Mar. 1997 (in Japanese).

[4] A. Kobayashi, T. Imai, A. Ando, E. Miyasaka, H. Akamatsu, S. Nakagawa, R. Oguro, K. Ozeki, S. Furui, J. Suzuki, and N. Shirai, “A Study on Continuous Speech Recognition System for Broadcast News,” Proc. Acoustical Society of Japan Fall Meeting, 3-1-9, Sept. 1997 (in Japanese).

[5] S. Furui, K. Takagi, A. Iwasaki, K. Ohtsuki, T. Matsuoka, and S. Matsunaga, “Japanese Broadcast News Transcription and Topic Detection,” Proc. DARPA Broadcast News Transcription and Understanding Workshop, Feb. 1998.

[6] B. Suhm and A. Waibel, “Towards Better Language Models for Spontaneous Speech,” Proc. ICSLP 94, S16-4, pp. 33 - 38, Sept. 1994.

[7] N. Kobayashi, Y. Nakano, Y. Wada and T. Kobayashi, “A Study on Word Unit Selection for LVCSR Using Entropy and Frequency of Phrase Observation,” IPSJ SIG Notes (SLP), Vol. 98, No. 12, Feb. 1998 (in Japanese)

[8] D. Klakow, X.Aubert, P. Beyerlein, R. Haeb-Umbach, M. Ullrich, A. Wendemuth, and P. Wilcox, “Language-Model Investigations Related to Broadcast News,” Proc. DARPA Broadcast News Transcription and Understanding Workshop, Feb. 1998.

[9] JUMAN (Japanese Morphological Analysis System) Version 3.0, http://pine.kuee.kyoto-u.ac.jp/ nl-resource/juman.html

[10] P. Clarkson, R. Rosenfeld, “Statistical Language Modeling Using the CMU-Cambridge Toolkit,” Proc. Eurospeech97, pp. 2707 - 2710, Sept. 1997.

[11] K. W. Church, P. Hanks, “Word Association Norms, Mutual Information and Lexicography,” Computational Linguistics, Vol. 16, No. 1, pp. 22 - 29, Mar. 1990.

[12] HTK: Hidden Markov Model Toolkit, Version 2.01, http://www.wntropic.com/htk/htk.html