GENERATION OF ADAPTIVE VOCABULARY LEXICON FOR JAPANESE LVCSR

Charles C.H. Jie

Philips Innovation Center, Taipei
24F, 66, Sec. 1 Chung-Hsiao W. Rd
Taipei, Taiwan, R.O.C
charles.jie@philips.com or chjie@ms23.url.com.tw

ABSTRACT

One of the thorniest problems of large vocabulary continuous speech recognition systems is the large number of out-of-vocabulary (OOV) words. This is especially the case for the languages like Japanese, which has many inflections, compound words and loanwords. The OOV words vary with the application domains. It's not realistic to have a big general-purpose lexicon including any possible OOV words. Furthermore, embedded speech recognition systems become more and more popular recently. They strongly demand an economical and effective exploitation of lexicon space. In this paper, we introduce a lexicon development model dealing with different kinds of OOV words with the help of linguistic morphological knowledge. It provides unsupervised, fast vocabulary adaptation among different application domains. In the experiments that adapt a lexicon of typical vocabulary, we were able to reduce the OOV rate by 40% and improve the word segmentation error rate by 27%. And the smaller the lexicon is, the more we benefit from the vocabulary adaptation.

1. INTRODUCTION

No sooner had we created our first lexicon for Japanese large vocabulary continuous speech recognition than we realized the challenge of out-of-vocabulary problems involved. Although 220 thousand words have been acquired from traditional dictionaries to build our lexicon, we can still find ten thousands of OOV words and one million occurrences of them in our corpora. The high OOV rate, ranging from 4.3% to 9.8%, may cause an impact on recognition system and account for a major share of recognition errors.

Inspecting to the text segmentation errors, several types of vocabulary problems can be identified. As an agglutinative language, the inflection of Japanese is very complicated and hard to handle. For example the gerund form 'natte' is one of the difficult problems for segmentation[1]. About compound and derived words, it's difficult to make new words recognizable without introducing new composing units (sub-words). We need to acquire prefixes and a suffixes from the derived words like '新-食糧-法'(new food law) to help identify new words composed in the similar way. Again we sometimes have to combine words to make a new entry in lexicon because the pronunciation of the combination can not be predicted by individual components, such as '一分'(one minute). We also found that Arabic numbers, alphabetic acronyms, and katakana[2] words are other prevailing types of OOV words appearing in news context like the Nikkei News[3]. However, not all of the OOV words should be added into the lexicon. We have to consider certain selection strategy to keep the lexicon efficient. An LVCSR system will have difficulty to perform well if it fails to cope with the vocabulary problems in early stages.

We consider an integrated approach to deal with the above issues. Motivated by Geutner's work[1] on similar problems and based on our needs, we propose a lexicon development model to easily and quickly

---

[1] To keep consistent with the segmentation of its plain form 'na-ru'(become), it should be segmented as 'na-tte'(te-form of naru). But it leads to a weird 'small-tsu'-initialized segment -tte.

[2] Katakana is one of Japanese writing symbol types. It's used to represent words of foreign origin.

[3] For example, the numbers of the three types of OOV words in half a year of Nikkei news are about 7000, 5000, and 44000, respectively.

generate the lexicon we want. It allows us to adapt the vocabulary of our lexicon from domain to domain, which may share only a quarter out of 55000 OOV words. It also enables us to freely reduce the lexicon size with minimum coverage loss.

2. DEVELOPMENT MODEL OF LEXICON

To fully automate and parameterize the building of our target lexicon, JSR (Japanese Speech Recognition) lexicon, a lexicon development model with related sources and targets has been specified, as shown in Fig.1. Samples of the lexicons are available in Fig. 2.

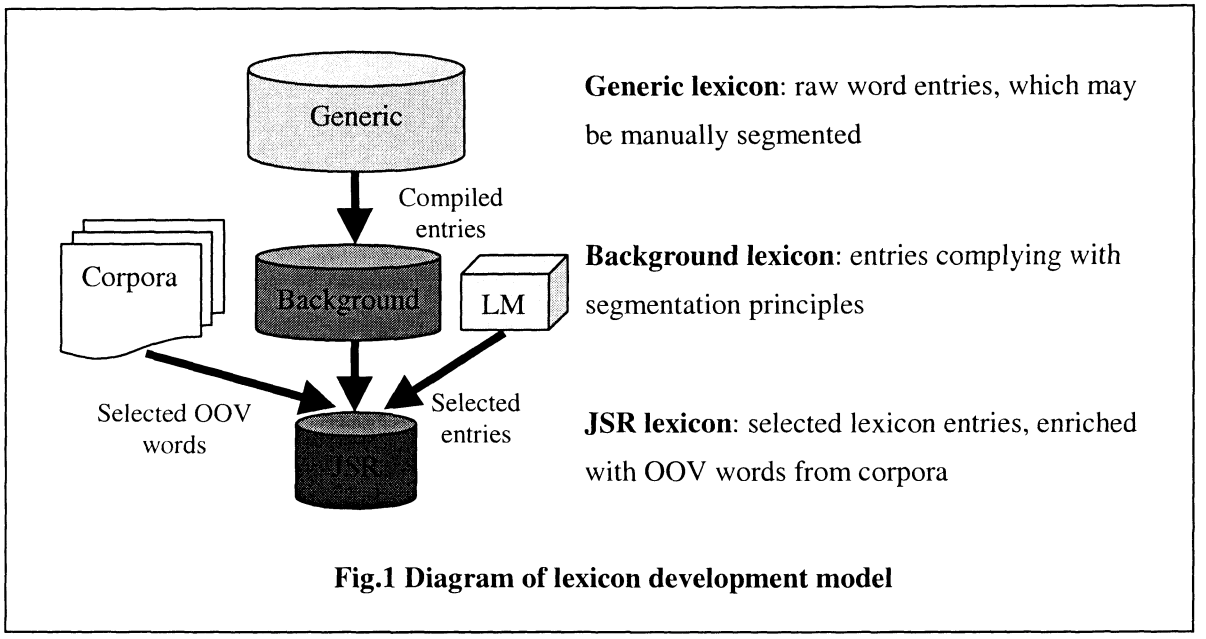

Fig.1 Diagram of lexicon development model

The **Generic Lexicon** contains our collection of vocabulary from various sources, mainly the word entries of traditional dictionaries available in hard copies. Manual segmentation has been done to make the word entries comply with our segmentation principles[2][4][5]. The segmentation for inflection words is especially important.

The **Background Lexicon** is compiled from generic lexicon and the segmentation information is referenced. For inflection words, only the stem is kept and a full list of inflection endings for verbs and adjectives[3] is added. Smaller segments within compound or derived words are collected because they can serve to reduce these types of OOV words.

The **JSR Lexicon** is a lexicon adapted from the basic vocabulary in background lexicon. It is usually enriched with the OOV words acquired from a specific corpus, which represents a domain we intend to adapt the lexicon to. The segmentation result and statistics of this corpus are taken

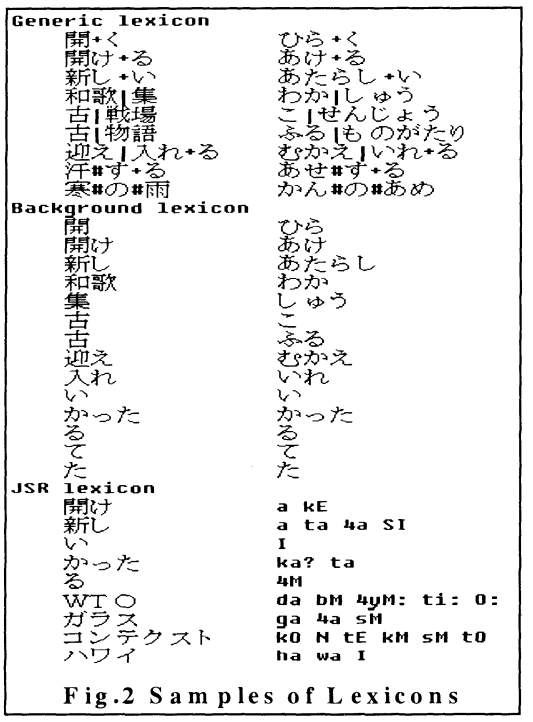

Fig.2 Samples of Lexicons

[4] Definitions of segmentation marks: '#' – word boundary, '|' – sub-word boundary, and '+' – boundary between the stem and ending of inflection words.

[5] Major differences from those of CallHome Japanese (http://www.ldc.upenn.edu/ldc/about/chjapanese.html) are (1) we decompose compounds and conventionalized expressions (2) we try to deal with auxiliaries and verbs/adjectives in a more systematic approach and lead to some new style of segmentation with much more details.

into account for the selection of effective words about the domain. A lexicon size is usually specified to meet various system requirements. In our evaluation, the lexicons of 64K, 32K, 16K and 8K size are built to test the effect of adaptation[6].

The pronunciation expression of the lexicons is usually in hiragana[7], while that of JSR lexicon is in SAMPA-Japanese[5]. Both the SAMPA transcription for hiragana and the pronunciation generation for OOV words are done by tools automatically.

Some basic facts about these lexicons are shown in Table 1:

Table 1 Entry number of Lexicons

	Word Number 1 (graphic form)	Word Number 2 (word + pronunciation)	Comments
Generic Lexicon	223K	243K	Typically 1/6 of JSR lexicon are OOV words from corpus in our experiments.
Background Lexicon	160K	181K	
JSR Lexicon	64K	77K	

The corpora used in our experiments are listed below:

Table 2 Corpora used for Training and Testing

	Contents	Size (Japanese characters)
Training 1	**Kyodo news**, half of a year Dated 1995/1/2-1995/6/30 (145 files)	18M
Training 2	**Nikkei news**, half of a year Dated 1993/12/1-1994/5/31 (181 files)	41M
Testing 1	**Kyodo news**, one day Dated 1994/11/30 (1 file)	43K
Testing 2	**Nikkei news**, one day Dated 1994/11/30 (1 file)	80K

The language model provides uni-grams that support corpus segmentation work and participate in the vocabulary selection of JSR lexicon. It is acquired from some other statistics of word counts over some corpus other than those used in our experiments.

The general quality of JSR lexicon is highly dependent on the segmentation principles and the consistency of the manual segmentation work over generic lexicon, which discussion is expected in other papers. The success of vocabulary adaptation among domains depends on the OOV word detecting module and the selection module. We will discuss them in next sections.

3. KEY MOUDULES

Fig. 3 illustrates the main part of the lexicon development system. The quoted numbers mark the results done by the following steps.

First, the "OOV word scanning" module detects OOV words from corpus with certain lexical rules. We create a temporary lexicon to include the vocabulary from background lexicon and the new OOV words from corpus. And at the same time, the word counts of OOV (acquired from the scanning result on corpus) and the word counts of the basic vocabulary (borrowed from a reference language model) are merged together for later use.

Second, the corpus is segmented with the support of the temporary lexicon. For each paragraph of the

---

[6] This number counts only the words of different graphic forms, while multiple pronunciations will give some extra entries.

[7] Hiragana is one of the writing symbol types of Japanese. It is used for most of the (sub-)words represented in pronunciation form. Kanji is another symbol type which is used to represent the words in graphic form.

text, word lattice is created to represent any possible words sequences within it. Dynamic programming is then applied to find the best path of segmentation with the help of the temporary lexicon.

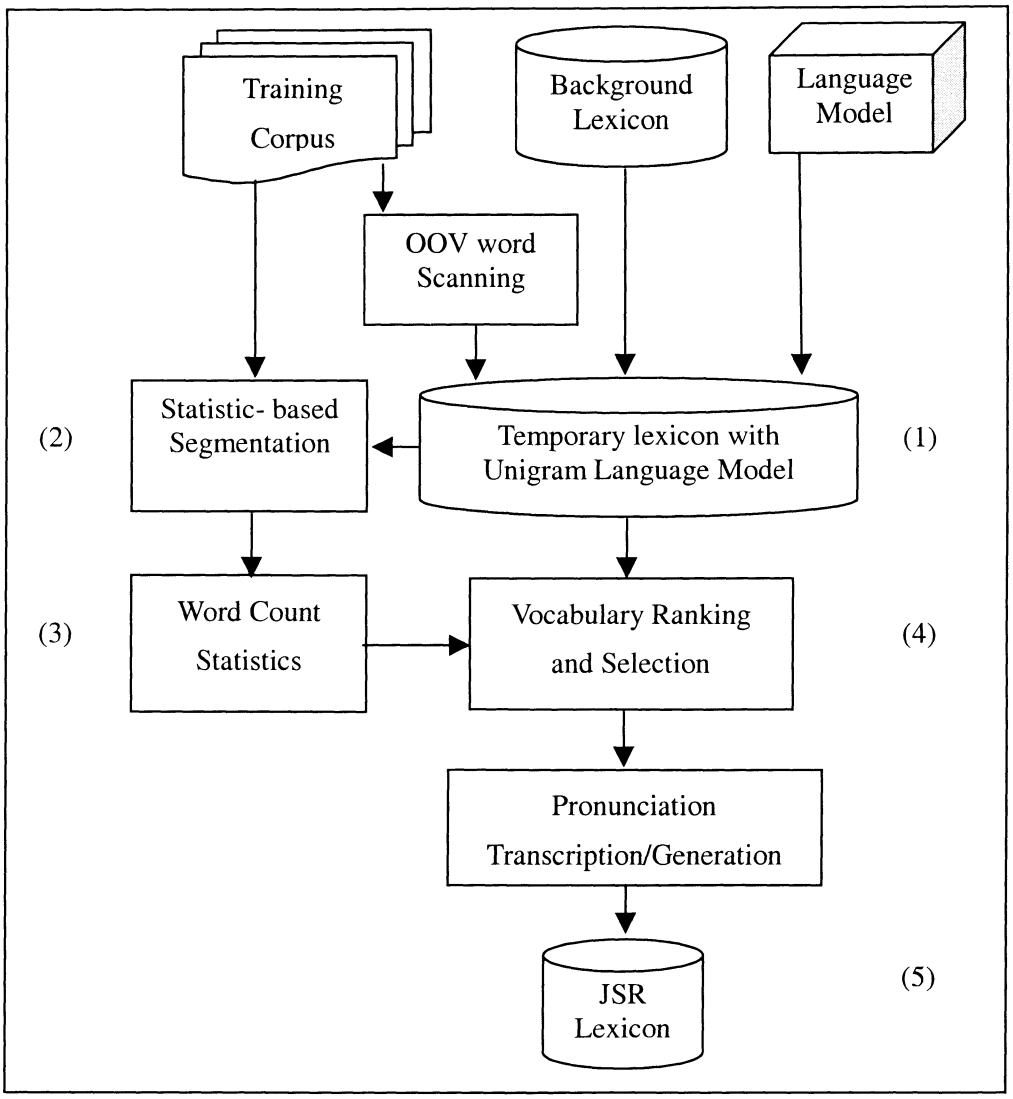

Fig.3 Main components of development framework

Third, the segmented result is passed to a statistical tool to get the word counts for the vocabulary in temporary lexicon. The word counts are considered as a new language model playing an important role in our vocabulary adaptation. Further OOV words can also be found in the statistics, but they take very limited percentage and can be ignored.

Fourth, considering only the vocabulary in temporary lexicon, we merge the new language model and the reference language model together with some weighting strategy. The sorting result of the hybrid language model is now ready to produce the target lexicon of any specified size.

Finally, in the wrap-up stage, we bind pronunciation information to these selected words. For the in-vocabulary words, we look up their pronunciations in background lexicon. For the OOV words, we generate their pronunciation based on the rules of thumb.

3.1 OOV WORDS SCANNING

We started from the analysis of OOV words in corpus. With a rough statistics, we found three types of OOV words prevail in the corpora. They are therefore chosen as our main targets and the following OOV statistics are basically referred to them. Table 3 shows the analysis of OOV rates on the corpora:

Table 3 OOV rates in Training Corpora

OOV rates	OOV / total by vocabulary number	%	OOV / total by word counts	%
Training corpus 1 (Kyodo news)	33.5K / 191K	18%	1065K / 21651K	4.9%
Training corpus 2 (Nikkei news)	55.5K / 211K	26%	1125K / 51844K	2.2%

The vocabulary OOV rate of Nikkei news is higher than Kyodo news, while the word OOV rate of Kyodo news is 2.2 times of that of Nikkei news. Kyodo news is a medium of general domain or multiple-purpose domain. Nikkei news is a professional economic medium. This suggests that there may exist different characters and problems for different domains or contexts.

We developed the regular expressions[8] in Fig. 4 to detect three types of OOV words in corpora:

- Katakana words – a katakana string which doesn't allow an initial prolongation mark (dash).
- Arabic numbers – a digit string with an optional decimal point part.
- Acronyms/English words – an alphabetic string.

```
OOV TYPE                    Regular Expression
-----------------------------------------------------
KATAKANA string:            [ ァ- ヲ][ー ァ- ヶ ]*

Arabic number:              [ 0 - 9 ]+( · [ 0 - 9 ]+)?

Acronym/English word:       [ A- Z a - z ]+
```

Fig 4. Patterns to scan OOV words

These patterns identify ten thousands of OOV words in corpora. In Table 4, we show their distribution about the three OOV types. It is interesting that though both corpora have larger vocabulary in katakana, the Kyodo has more words in Arabic numbers. This causes our curiosity to study the common OOV words between the two corpora. In Table 5 we see the limited common vocabulary takes a major part in word counts.

Table 4 Distribution of OOV words over three OOV types

Corpus	OOV type	OOV / total by vocabulary number	%	OOV / total by word counts	%
Training corpus 1 (Kyodo news)	KATAKANA	18.8K / 33.5K	56%	410K / 1065K	38%
	Arabic number	12.6K / 33.5K	38%	583K / 1065K	55%
	Acronym/English	2.1K / 33.5K	6%	72K / 1065K	7%
Training corpus 2 (Nikkei news)	KATAKANA	43.3K / 55.5K	78%	814K / 1125K	72%
	Arabic number	7.5K / 55.5K	13%	209K / 1125K	19%
	Acronym/English	4.8K / 55.5K	9%	103K / 1125K	9%

[8] There exist various flavors of Regular Expressions[4], such as *Perl, grep, vi, emacs, sed, awk,* etc. Our patterns are written in *Perl* flavor. And note that double-byte capacity of *Perl* is assumed, otherwise they need to be transcribed.

Table 5 Rates of Common OOV words

Rate of Common OOV	For OOV type	Common / All by vocabulary number	%	Common / All by word counts	%
Common part / Training corpus 1 (Kyodo news)	KATAKANA	10.2K / 18.8K	54%	379K / 410K	92%
	Arabic number	4.0K / 12.6K	32%	559K / 583K	96%
	Acronym/English	1.2K / 2.1K	57%	64K / 72K	89%
	Total	15.3K / 33.5K	46%	1002K / 1065K	94%
Common part / Training corpus 2 (Nikkei news)	KATAKANA	10.2K / 43.3K	24%	711K / 814K	87%
	Arabic number	4.0K / 7.5K	53%	204K / 209K	98%
	Acronym/English	1.2K / 4.8K	25%	88K / 103K	86%
	Total	15.3K / 55.5K	28%	1003K / 1125K	89%

The above statistics are based on the assumption that we don't have those "OOV words" in our background lexicon. But indeed some of the katakana "OOV words" can be found in our fundamental vocabulary. As you can see in Table 6, about one tenth of the OOV in corpora already exists in our collection of basic vocabulary. There remains a half to three quarters of the OOV occurrences waiting for our solution.

Table 6 The "OOV words" available in Background Lexicon

In-voc Rate of the "OOV words"	In-voc / All by vocabulary number	%	In-voc / All by word counts	%
Training corpus 1 (Kyodo news)	3.2K / 33.5K	9.5%	272K / 1065K	26%
Training corpus 2 (Nikkei news)	5.0K / 55.5K	9.0%	546K / 1125K	48%

3.2 VOCABULARY RANKING AND SELECTION

In order to select the most effective words to build the target lexicon, we rank the in-vocabulary and out-of-vocabulary words according to their exploitation in corpora. We take the word counts of CallHome Japanese corpus as basic reference data, which are stored as our Language Model. The word counts acquired from the training corpora are even important sources in the sense of adaptation. In our current experiments, we only sum up the word counts of correspondent vocabulary to have a new ranking, which represents the exploitation in both resources. However, when we have further word counts from some other corpora, we may need to consider certain kind of normalization or weighting strategy among the different sources

There are two exceptions in our vocabulary selection based on word counts:

- The Arabic numbers are excluded, though we prepare and keep them in temporary lexicon. We need them in temporary lexicon because they help the corpus segmentation. However we don't need them in target lexicon because they can be treated as a sequence of Arabic digits or Kanji number segments[9], depending on how they are pronounced in speech recognition.
- Some inevitable segments are always included even if they are never trained or seldom seen in corpora. Such words include the inflection endings, punctuation marks, symbols etc.

[9] Such as sanzen(three thousand) and ropiaku(six hundred). They are kept in this way (number-unit pair) because of the complicated pronunciation.

4. PERFORMANCE EVALUATION

4.1 EVALUATION MODEL

The above development model provides us a platform to generate specific lexicons easily and quickly. In principle we have full freedom to specify the corpus to adapt to and the lexicon size meeting the system requirements. We thus make the following plan to evaluate (1) the effect of adaptation to various domains and (2) the influence of reducing lexicon size.

We build the baseline JSR lexicon with the vocabulary selection based on the word counts[10] of CallHome Japanese corpus by Linguistic Data Consortium. We then build the adapted lexicons for Kyodo News domain, Nikkei News domain, and both. These lexicons are all evaluated for their coverage rate and segmentation error rate over our testing data, including Kyodo and Nikkei News both. All these evaluations are done for lexicons of size 64K, 32K, 16K and 8K.

Both the testing files, kyodo941130 (25.7K words) and nikkei941130 (164K words), are kept separated from the training materials and involved language models. They are segmented by manual work according to our segmentation principles.

The coverage rate (= in-vocabulary words / total) is evaluated by counting the words in manual segmented test file against the vocabulary of a specific lexicon. The segmentation error rate (= (substitution + insertion + deletion) / total words) is evaluated by comparing the reference (manually segmented) data with the automatic segmentation result, which is done by a statistical approach with a lexicon involved.

We acquire (4 lexicons) x (4 adaptations) x (2 testing domains) x (2 kinds of rates) = 64 major results in the experiments and we'll discuss them in the following sections.

4.2 RESULTS OF COVERAGE RATE EVALUATION

The coverage rate is a very straightforward index to the performance of lexicons. Table 7 records the coverage rate on the Kyodo News test material for the baseline lexicons and adapted lexicons, while Table 8 on the Nikkei News test material. The following facts can be observed:

Table 7 Coverage rates tested on Kyodo News

Coverage rate	8K	16K	32K	64K
Baseline	51.91	54.43	64.72	87.83
Adapted to Kyodo	90.59	93.04	93.73	93.94
Adapted to Nikkei	89.6	92.18	93.4	93.81
Adapted to Both	90.18	92.83	93.73	93.97

Table 8 Coverage rates tested on Nikkei News

Coverage rate	8K	16K	32K	64K
Baseline	50.13	52.95	64.46	87.74
Adapted to Kyodo	86.51	90.27	92.05	92.92
Adapted to Nikkei	88.75	91.87	93	93.34
Adapted to Both	88.67	91.94	93.02	93.36

- The coverage rate can be significantly improved from 88% to 93% for 64K lexicon if the lexicon is adapted. It improves even more for smaller lexicons. It shows that the OOV words discovered in training materials are really useful in the domains of our daily life. More than 40% of the remaining OOV problems of baseline lexicon can be resolved with our adaptation.
- The coverage rate drops very fast, from 88% to 52%, for baseline lexicon if we cut the lexicon size from 64K down to 8K, while it drops not much for adapted lexicons. The vocabulary selection for

[10] The word counts acquired from CallHome Japanese lexicon sum up to 748K.

baseline lexicons is only based on the word counts of CallHome Japanese corpus, which is quite small compared to our training materials. This indicates sufficient training material for domain adaptation is important.

- If adapted to Kyodo News, we'll have better coverage rate for Kyodo News than for Nikkei News. The adaptation also takes effect for Nikkei News. Checking the middle lines of Table 7 and 8, this adaptation effect becomes more and more evident when we cut the lexicon size down. We found that training with Kyodo can still have a pretty good coverage on Nikkei, and vice versa. Maybe it's because both domains are related to daily news and both training materials are huge in amount. Some more experiments can be conducted in this aspect.
- If we adapt the lexicon with the training materials of both domains, we'll have the best or near the best coverage rate. It's true on both testing materials. The additional training with other domains may introduce useful missing vocabulary, but it may also compete with the vocabulary from the training with target domain.

We illustrate the above tables in Fig 5 and Fig 6, which better depict the trend of coverage rate for various sizes of lexicon. Since the coverage rate drops slowly for a well-adapted lexicon, it looks the embedded system developers may feel free to cut the lexicon size down without much loss in this range.

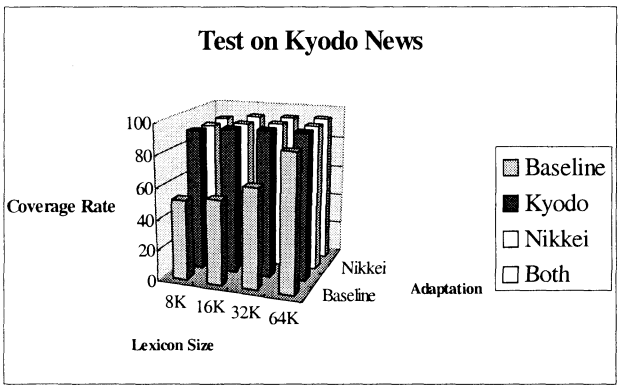

Fig. 5 Coverage rates on Kyodo News

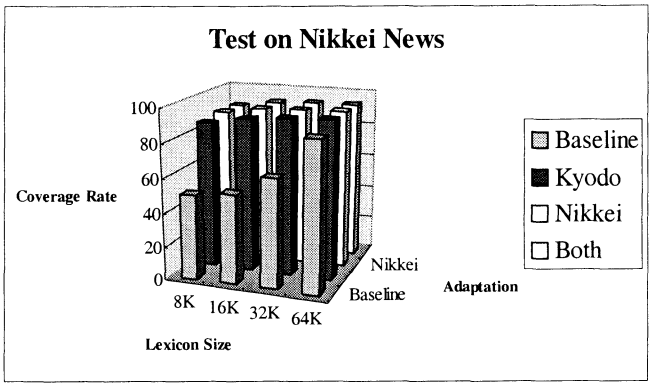

Fig.6 Coverage rates on Nikkei News

In addition to the above aspect of word coverage rate evaluation, people may be interested in another kind of coverage rate – the vocabulary coverage rate. For a testing file, this rate (= in-vocabulary / whole-vocabulary) refers to the ratio of the vocabulary available in lexicon and the total vocabulary of the file. We list these lexicons' vocabulary coverage rates in Table 9 and 10 for reference.

Table 9 Vocabulary coverage rates tested on Kyodo News

Voc. coverage rate	8K	16K	32K	64K
Baseline	9.32	14.13	30.39	64.75
Adapted to Kyodo	63.58	73.91	77.09	78.00
Adapted to Nikkei	60.59	71.46	75.91	77.76
Adapted to Both	62.03	73.19	77.01	78.24

Table 10 Vocabulary coverage rates tested on Nikkei News

Voc. coverage rate	8K	16K	32K	64K
Baseline	6.23	12.92	26.03	54.55
Adapted to Kyodo	34.87	53.02	62.73	67.06
Adapted to Nikkei	38.08	57.38	66.22	69.34
Adapted to Both	37.85	57.59	66.30	69.53

Although the word coverage rates on the two test files are similar, the vocabulary coverage rate is much higher on Kyodo News. This reveals that the Nikkei text has a heavier load on a smaller vocabulary.

4.3 RESULTS OF ERROR RATE EVALUATION

We take a simpler approach here to evaluate the function of our lexicons designed for speech recognition. The performance of statistics-based word segmentation is also highly dependent on the lexicon involved. We check the segmentation result against the manually segmented data. The occurrences of substitution, insertion and deletion are summed up and divided by the total word number to give the segmentation error rate.

Inevitably certain language model has to participate in the segmentation process. We simply adopt unigrams here, which are word counts based on the training materials. For baseline lexicon, only CallHome Japanese's word counts are involved. For the lexicon adapted to Kyodo News, the word counts of Kyodo training data are added. The same language model adaptation is also done for Nikkei part. For the lexicon adapted to both Kyodo and Nikkei, both the training results on Kyodo and Nikkei are added to the basic word counts from CallHome corpus. The inclusion of word counts from correspondent training data is to guarantee the new vocabulary of adapted lexicon can work in probability-based segmentation.

The results of error rate evaluation are listed in Table 11 and 12. The following points can be observed:

Table 11 Error rates tested on Kyodo News

Error rate	8K	16K	32K	64K
Baseline	46.24	44.19	47.88	38.06
Adapted to Kyodo	33.17	29.36	28.08	27.57
Adapted to Nikkei	35.52	31.46	28.99	27.98
Adapted to Both	33.9	29.93	28.23	27.63

Table 12 Error rates tested on Nikkei News

Error rate	8K	16K	32K	64K
Baseline	44.78	41.25	45.03	37.12
Adapted to Kyodo	40.71	35.21	32.28	31.03
Adapted to Nikkei	36.96	32.8	31.04	30.44
Adapted to Both	37.05	32.65	31.01	30.42

- The error rate is improved remarkably from 38% to 27.5% for the 64K lexicon if the lexicon is adapted to Kyodo and tested on Kyodo. It's about 27% of improvement. In Nikkei's case, the error rate for adapted lexicons is generally higher and its improvement by adaptation is not as obvious as Kyodo. This implies certain difference in character between the two news corpora.
- The adaptation to either news domain does reduce the error rate better for the domain in question. This is especially obvious with smaller lexicons. This trend can be checked in the middle rows of the tables and is clearer depicted with Fig. 7 and 8. Note lexicons are arranged in reverse order than the diagrams of coverage rate.
- Adaptation to both domains acquires as good result as the adaptation dedicating to a single domain. In fact, it makes the best total performance.

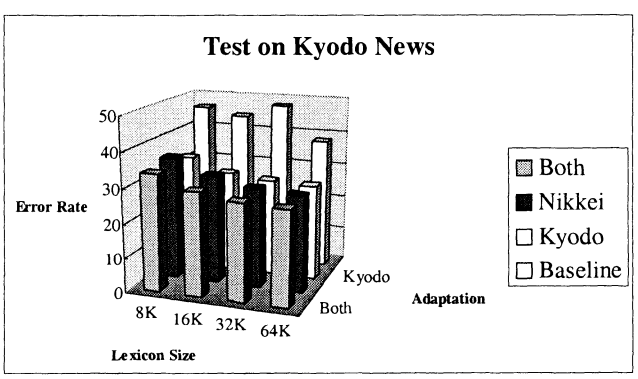

Fig. 7 Error rates on Kyodo News

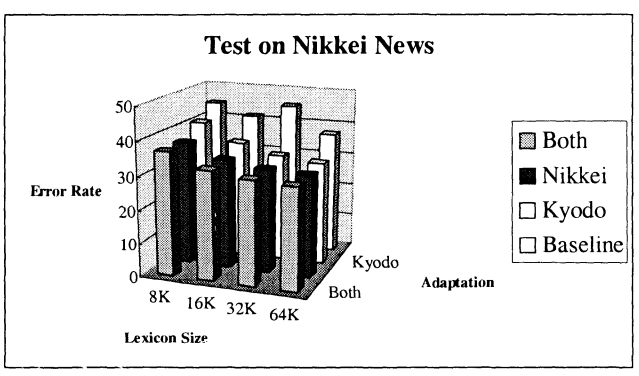

Fig. 8 Error rates on Nikkei News

A remark for our experiments: Because we don't include Arabic numbers in the lexicons and don't apply any mechanism to deal with them in segmentation process, the above coverage rates and error rates indeed include the influence from it, especially the latter. If we remove Arabic numbers from testing data, the coverage rates will go up about 1% generally, and the error rates can go down about 3% for Kyodo or less than 1% for Nikkei.

5. CONCLUSIONS

This paper describes a lexicon development model and how it improves our lexicon for speech recognition. The vocabulary adaptation selects effective words from basic vocabulary and supplements it with domain-specific OOV words matching certain patterns. It achieves remarkable improvement in the evaluation for word coverage rate and segmentation error rate. And the experiment results also provide valuable information for developing a compact lexicon, which is required in embedded speech recognition system.

We are developing further experiments to evaluate the effect of vocabulary adaptation in a speech recognition system, which is more complicated. The interaction among lexicon, language model and acoustic component is worth exploring. We may find new problems and improve our lexicon development model accordingly.

6. ACKNOWLEDGEMENTS

I'd like to thank my colleague Yumiko Sasaki (linguistic specialist) and my old friend Professor Zhao-Ming Gao (in National Taiwan University) that help review my paper and provide valuable suggestions. I also appreciate Dr. Detlev langmann and Dr. Paul Lin. Without their support and encouragement, this paper would not be possible. And I'd lik to give the last 'thank you' to my wife for the reason well known.

7. REFERENCES

[1] P. Geutner, M. Finke and P. Scheytt, "Adaptive Vocabularies for Transcribing Multilingual Broadcast News", ICASSP98

[2] Detlev Langmann, Sachiko Morishita, and Charles Jie, "Design Specification Guidelines for Japanese Word Segmentation", Philips Innovation Center, Taipei and Philips Speech Processing, 1999.

[3] Sachiko Morishita and Detlev Langmann, "Specification of Japanese Infection Words and Tree Diagrams", Philips Speech Processing, 1999.

[4] Detlev Langmann, Y.C. Chu, Joseph Huang, Yumiko Sasaki and Sachiko Morishita, "SAMPA-J - Specification of Japanese SAMPA", Philips Innovation Center, Taipei, 1999.

[5] Jeffrey E. F. Friedl, "Mastering Regular Expressions", O'REILLY, 1998