UNSUPERVISED ADAPTATION OF A STOCHASTIC LANGUAGE MODEL USING A JAPANESE RAW CORPUS

Gakuto KURATA, Shinsuke MORI, Masafumi NISHIMURA

IBM Research, Tokyo Research Laboratory, IBM Japan, Ltd.
1623-14 Shimotsuruma Yamato-shi Kanagawa, 242-8502, Japan
{gakuto,forest,nisimura}@jp.ibm.com

ABSTRACT

The target uses of Large Vocabulary Continuous Speech Recognition (LVCSR) systems are spreading. It takes a lot of time to build a good LVCSR system specialized for the target domain because experts need to manually segment the corpus of the target domain, which is a labor-intensive task. In this paper, we propose a new method to adapt an LVCSR system to a new domain. In our method, we stochastically segment a Japanese raw corpus of the target domain. Then a domain-specific Language Model (LM) is built based on this corpus. All of the domain-specific words can be added to the lexicon for LVCSR. Most importantly, the proposed method is fully automatic. Therefore, we can reduce the time for introducing an LVCSR system drastically. In addition, the proposed method yielded a comparable or even superior performance to use of expensive manual segmentation.

1. INTRODUCTION

Recently Large Vocabulary Continuous Speech Recognition (LVCSR) systems are able to recognize speech in a general domain with high accuracy. This motivates us to apply LVCSR systems to various domains such as call centers, court reports, medical reports, university lectures, and so on [1].

Since a speech in a specific domain contains many domain-specific words and expressions, it is difficult for a general LVCSR system to recognize speech in a specific domain. Considering that domain-specific words are likely to characterize their domain, misrecognition of these words causes a severe quality degradation of the LVCSR application. In addition, misrecognition of these words causes misrecognition of surrounding words [2].

In order to apply LVCSR system to a specific domain, it is necessary to add domain-specific words to the lexicon for LVCSR and build a domain-specific Language Model (LM). Using the corpus of the target domain is effective, according to related studies [3, 4]. Fortunately, a lot of articles are computerized these days and we can easily get a corpus of the target domain. As is well known, in Japanese, like other Asian languages, no spaces exist between words [5]. Therefore, it has been necessary to segment the target domain's corpus into words.

The ideal method is as follows: (1) Experts manually segment a corpus of the target domain. (2) Domain-specific words that only appear in this corpus are added to the lexicon for LVCSR. (3) The domain-specific LM is built from this correctly segmented corpus. However, in this method, every time the target domain changes, experts need to manually segment a corpus of the new target domain. This is not realistic, considering that the target of LVCSR should be unconstrained. In order to adapt LVCSR to various domains, a fully automatic method is necessary.

An automatic word segmenter is available to segment the corpus [6]. However, segmentation errors inevitably occur. In particular, domain-specific words are likely to be analyzed wrongly because an automatic word segmenter is not trained with the domain-specific corpus. Considering this, it has been difficult to fully automatically adapt an LVCSR system to a specific domain.

In this paper, we propose a fully automatic method to adapt an LVCSR system to a specific domain. In this method, a Japanese corpus that is not segmented into words (a raw corpus) is regarded as stochastically segmented. We build a domain-specific LM from a raw corpus of the target domain. In addition, all character strings in the raw corpus can be treated as words. Therefore, domain-specific words can be regarded as words, added to the lexicon, and assigned proper probabilities based on their lexical contexts. The details are described in Sec. 2.

Experiments showed that an LVCSR system applied fully automatically with the proposed method achieved comparable and even superior performance to an LVCSR system created expensively using experts' manual segmentation.

2. PROPOSED METHOD

In this section, we describe a new method to adapt an LVCSR system to a new domain using a raw corpus of the target domain. This method is fully automatic and therefore, not expensive and time-consuming. This method has three stages:

1. Segment the raw corpus stochastically.
2. Build a word n-gram model from the stochastically segmented corpus.
3. Add probable words into the lexicon for LVCSR.

At the end of this section, we summarize the proposed method.

2.1. Stochastic Segmentation

As already mentioned, in Japanese sentences, all of the words are concatenated and there is no word boundary information. In order to build an LM, it has been necessary to deterministically judge whether or not the character boundary is a word boundary. We call a corpus that is segmented into words deterministically a "Deterministically Segmented Corpus". An example is shown in Fig. 1.

In contrast to the deterministic segmentation, stochastic segmentation was proposed [7]. In this method, an unsegmented raw corpus of n_r characters is regarded as a sequence of characters $\boldsymbol{x} = x_1 x_2 \cdots x_{n_r}$. Then the probability p_i that a word boundary exists

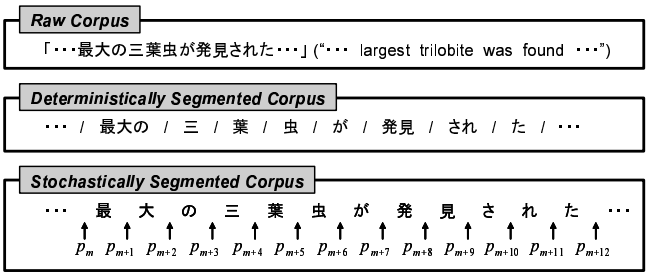

Fig. 1. Examples from a Segmented Corpus

after the i-th character x_i for each $i \in \{1, 2, \cdots, n_r - 1\}$ is calculated. We call a corpus that is annotated with these word boundary probabilities (p_i) a “Stochastically Segmented Corpus”. An example is also shown in Fig. 1.

2.2. Word n-gram Model from the Stochastically Segmented Corpus

Given a stochastically segmented corpus of n_r characters annotated with word boundary probabilities p_i, the number of words in the corpus (stochastic word zero-gram) is calculated as follows:

$$f_r(\cdot) = 1 + \sum_{i=1}^{n_r - 1} p_i \quad .$$

A character sequence $\boldsymbol{x}_{i+1}^{i+k}$ in the raw corpus is treated as a word $w = \boldsymbol{x}_{i+1}^{i+k}$ if and only if there is a word boundary before and after the sequence and there is no word boundary inside the sequence. Thus, the stochastic uni-gram frequency f_r of a word w in the raw corpus is defined by the summation of the stochastic frequencies at all occurrences O_1 of the character sequence of the word w over all of the character sequences in the raw corpus as follows:

$$f_r(w) = \sum_{i \in O_1} p_i \left[\prod_{j=1}^{k-1} (1 - p_{i+j}) \right] p_{i+k} \quad ,$$
$$\text{where} \quad O_1 = \{ i \,|\, \boldsymbol{x}_{i+1}^{i+k} = w \} \quad .$$

A word uni-gram probability is obtained by dividing the stochastic uni-gram frequency of the word by the stochastic word zero-gram frequency.

$$P_{\text{1-g}}(w) = f_r(w) \,/\, f_r(\cdot) \quad .$$

Similar to the word uni-gram probability, the word n-gram probability is obtained by dividing the stochastic n-gram frequency of the word sequence by the stochastic $(n-1)$-gram frequency.

2.3. Probable Character Strings Added to the Lexicon

Using the stochastic segmentation, all of the character strings appearing in the domain-specific corpus can be treated as words. Therefore, Out-Of-Vocabulary (OOV) words which are not included in the general lexicon can be regarded as words. However, a lot of meaningless character strings are also included in these possible words. We used a traditional character-based approach to judge whether or not a character string is appropriate as a word [8, 9]. This approach is based on the frequencies of the character strings in the corpus[10]. Only the character strings regarded as an appropriate word are added to the lexicon for LVCSR. Unfortunately, the long history of related research shows that detecting words from a Japanese text is still difficult. As a result, a lot of meaningless character strings may remain as words, resulting in a large number of added words. The meaningless character strings added to the lexicon may have a negative influence.

2.4. Summary of the Proposed Method

As regards time, the proposed method has an advantage, because it only requires a raw corpus and doesn't need labor-intensive manual segmentation to adapt an LVCSR system to the target domain.

From the aspect of performance, OOV words can be treated as words. In addition, proper n-gram probabilities are assigned to OOV words and the word sequences containing OOV words. Theoretically speaking, this contributes to the performance of LVCSR. We conducted experiments to assess the advantages and disadvantages.

3. BASIC MATERIAL

In this section, we will briefly explain the acoustic model, the general LM, and the general lexicon used in common in the experiments described in the next section.

3.1. Acoustic Model

We used a spontaneous speech corpus of 83 hours long to train the acoustic model (AM). Phones are represented as context-dependent, three-state, left-to-right HMMs. The HMM states are clustered using a phonetic decision tree and the number of leaves was 2,728. Each state of the HMMs is modeled using a mixture of Gaussians, and the number of mixtures was 11.

3.2. General LM and General Lexicon

We have a large corpus of a general domain. This corpus is mainly composed of newspaper articles. A small part of the corpus was segmented into words by experts. The rest was segmented automatically by the automatic word segmenter and roughly checked by experts. We built from this corpus a general LM and a general lexicon which were used in common in all of the experiments. The number of words in the general corpus was 24,442,503. The general lexicon contained 45,402 unique words.

We used word bi-gram models instead of tri-gram models because of the empirical results and the computational requirements. Our pilot experiments didn't show a significant difference between bi-grams and tri-grams, though the computational costs were significantly different.

4. EXPERIMENTS

We conducted the experiments on lectures of *the University of the Air*. *The University of the Air* delivers broadcast lectures via TV and radio. The content of the lectures is specialized. Domain-specific words which never appear in newspaper articles are often used. For each lecture, we built LVCSR systems specialized for that lecture.

We selected three lectures for the experiments. The subjects and the sizes of each lecture speech are shown in Table 1. The OOV rates based on the general lexicon are also shown in Table 1.

For each lecture, we prepared related raw corpora which corresponded to each lecture. These related corpora are mainly composed of the textbooks which are published by *the University of the Air*. Table 2 shows the sizes of the related corpora.

Table 1. Overview of Lecture Speeches

Lecture ID	Subject	Total # of Words	OOV rate
$\mathcal{B}$	Biology	2,260	5.7%
$\mathcal{M}$	Music	2,679	4.5%
$\mathcal{G}$	Geoscience	2,270	6.3%

Table 2. Two Sizes of Related Corpora (Total # of Characters)

Lecture ID	Small	Large
$\mathcal{B}$	10,641	73,437
$\mathcal{M}$	16,251	88,996
$\mathcal{G}$	10,892	69,617

To examine the effect of the sizes, we prepared small and large related corpora for each lecture, as shown in Table 2. The small corpus is approximately equivalent to 20 pages of the textbook. The large corpus is approximately equivalent to one entire textbook. The small corpus is a subset of the large corpus. Fig. 2 shows the flow of the experiments. As shown in the gray box in Fig. 2, we built the LM for each lecture from the corpus related to that lecture. We compared three methods, namely the *Ideal* method, the *Automatic* method, and the *Proposed* method. The *Ideal* method is based on the manual segmentation and considered to be the best option to maximize the performance of LVCSR. However, it is time-consuming to manually segment the related corpus. In contrast, the *Proposed* method is a fully automatic method and saves a lot of time. For comparison, as an existing fully automatic method, we did experiments using the automatic word segmenter.

We describe the details of these three methods as follows:

Ideal The experts segmented the related raw corpus manually. The LM was built from this correctly segmented corpus. All of the OOV words which only appear in this correctly segmented corpus were added to the general lexicon for LVCSR.

Automatic The automatic word segmenter segmented the related raw corpus. The LM was built from this segmented corpus. All of the OOV words appearing in this segmented corpus were added to the general lexicon for LVCSR.

Proposed The raw corpus was stochastically segmented as described in Sec. 2.1. The LMs were built using the method in Sec. 2.2. The probable character strings were added to the general lexicon using the method in Sec. 2.3.

After we built the domain-specific LM from the related corpus with one of these three methods, we interpolated it with the general LM. We used this interpolated LM and the common AM for LVCSR. For comparison, we also made an experiment using only the general LM and the common AM.

5. EVALUATION

In this section, we explain the results and discuss them.

5.1. Results and Discussion

We compared the recognition accuracy of the three methods described in Sec. 4. To measure the recognition accuracy, we used the Character Error Ratio (CER). The reason is that in Japanese ambiguity exists in word segmentation. For example, "Governor of Tokyo (東京都知事)" can be segmented into words in four ways: (1) "東京都知事", (2) "東京都 / 知事", (3) "東京 / 都知事", and (4) "東京 / 都 / 知事". In all cases, the same characters are used

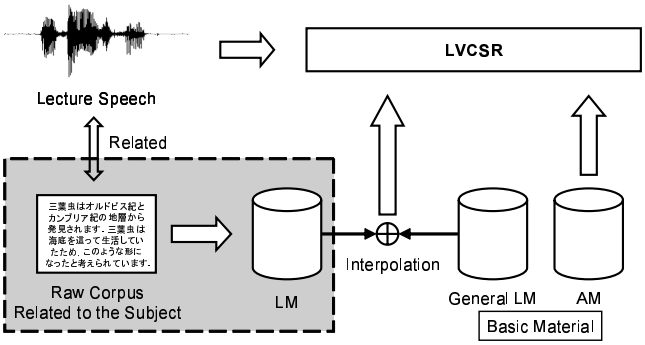

Fig. 2. Overview of the Experiments

and the number of the characters remains 5. However, the number of the words seems to change from 1 to 3 because of the ambiguity. For domain-specific words, this ambiguity is likely to increase. Therefore, the CER is suitable for the criterion in these experiments. In addition, we estimated WER based on the CER and the average number of characters $\bar{n}$ per one word. We named this criterion "*e*WER" and this was defined as follows : $e\text{WER} = (1-(1-\text{CER})^{\bar{n}}) \times 100$.

Table 3 shows the CERs. The values in parentheses are the *e*WERs. The values in square brackets are the numbers of words added to the general lexicon for LVCSR. The second column from the left shows the CERs when only the general LM was used. This is to say that the related corpus was not used. The other columns are the CERs when the interpolated LMs were used.

For example, the cell in the bottom-right corner has the following meanings: (1) The domain-specific LM was built with the *Proposed* method from the large raw corpus related to the lecture $\mathcal{G}$. (2) The interpolated LM of the general LM and the domain-specific LM for the lecture $\mathcal{G}$ was used for LVCSR. (3) 3,947 words were added to the general lexicon for LVCSR. (4) The result of LVCSR for the lecture $\mathcal{G}$ was a 22.9% CER and a 46.4% *e*WER.

We explain the results and compare the three methods.

5.1.1. Without a Related Corpus

Looking at the second column, the performance was not satisfactory when the related corpus was not used, as we had anticipated. The reason for this is that the general LM and the general lexicon were based on a general domain.

5.1.2. With a Small Related Corpus

The third, fourth, and fifth columns show the performances of LVCSR using the small related raw corpora.

Comparing the second column with the third, fourth, and fifth columns, all of the LMs using the related corpora improved the performances. Even though the sizes of the related corpora are small, they contribute to an improvement of LVCSR in the specific domain, as a previous study had reported [4]. For example, in the lecture $\mathcal{B}$, the CER was 27.0% when only the general LM was used. It was decreased to about 12%.

Looking at the third, fourth, and fifth columns, the *Ideal* method showed the best performance as we had expected. These results are reasonable.

Comparing the fourth and fifth columns, the *Proposed* method decreased the CER more than the *Automatic* method. Note that all processing of the *Automatic* and the *Proposed* methods are completed fully automatically. This means that the *Proposed* method

Table 3. CER (*e*WER) [%]

Lecture ID	General LM Only	Size of Raw Corpora / *Methods for Adaptation*					
		Small (about 20 pages)			Large (about one whole textbook)		
		Ideal	*Automatic*	*Proposed*	*Ideal*	*Automatic*	*Proposed*
$\mathcal{B}$	27.0 (53.0)	11.5 (25.4) [+ 271 words]	12.9 (28.2) [+ 310 words]	**11.6 (25.6)** [+ 1,315 words]	N/A	13.7 (29.8) [+ 1,538 words]	**11.0 (24.4)** [+ 4,705 words]
$\mathcal{M}$	24.9 (49.7)	17.3 (36.6) [+ 287 words]	17.9 (37.7) [+ 215 words]	**17.6 (37.2)** [+ 722 words]	N/A	18.0 (37.9) [+ 1,417 words]	**17.0 (36.1)** [+ 4,698 words]
$\mathcal{G}$	28.0 (54.5)	23.3 (47.1) [+ 224 words]	25.0 (49.9) [+ 280 words]	**23.1 (46.8)** [+ 832 words]	N/A	23.4 (47.3) [+ 1,050 words]	**22.9 (46.4)** [+ 3,947 words]

⋆ The values in square brackets are the numbers of words added to the general lexicon for LVCSR.

achieved a further improvement over the *Automatic* method without increasing the time. In the lecture $\mathcal{B}$, the CER decreased from 12.9% to 11.6%. In the lecture $\mathcal{G}$, the CER also decreased from 25.0% to 23.1%. These are equivalent to about 10% error reduction.

Most importantly, comparing the third and fifth columns, the performance of the *Proposed* method is close to that of the *Ideal* method. The *Proposed* method showed comparable performance to the *Ideal* method, simultaneously saving a lot of time.

5.1.3. With a Large Related Corpus

The sixth, seventh, and eighth columns show the performances of LVCSR using the large related raw corpora. The sixth column is N/A because the *Ideal* method cannot be used with large raw corpora. It is not realistic to manually segment the whole textbook.

In order to examine the effect of the size of corpora when the *Proposed* method is used, we compare the fifth and eighth columns. The LMs based on the large corpora always worked better than the LMs based on the small corpora. Considering this result, the larger the raw corpora are, the better the performances of LVCSR are with the *Proposed* method.

Then the eighth column is compared with the third column (small corpora with the *Ideal* method) which was considered to be the best option for the realistic conditions. For all of the lectures, the *Proposed* method using large raw corpora yielded better performances with much less time than the *Ideal* method manually using small raw corpora. In the case of lecture $\mathcal{B}$, the CER was reduced from 11.5% to 11.0%. This means that the *Proposed* method showed the best performance in a fully automatic way.

5.1.4. Number of Words Added to the Lexicon

The numbers of words added to the lexicon are larger when the *Proposed* method is used as we had anticipated in Sec. 2.3. A lot of character strings inappropriate as words are included in these words. However, our proposed method assigns very small probabilities to the meaningless character sequences. The accuracy of LVCSR shows that there was no negative influence.

5.1.5. Summary

From the observations above, we achieved the best performance using the fully automatic *Proposed* method with the large raw corpus.

Since a lot of articles are computerized these days, it is not a difficult task to collect large raw corpora. In contrast, it is and will be an expensive and time-consuming task to manually segment a raw corpus. Therefore, our proposed method which only requires a raw corpus is practical. The results of our experiment show that just collecting a relevant corpus improves the performance of LVCSR more than expensively segmenting the corpus. This result is promising in introducing LVCSR into various new domains.

6. CONCLUSION

An LVCSR system built for a general domain is not good at recognizing speeches in a specific domain. In order to apply an LVCSR system to a new specific domain, it has been necessary to prepare a corpus of the target domain, manually segment it into words, and build an LM. This was the ideal method to maximize the performance of LVCSR, but needed labor-intensive segmentation. In this paper, we propose a new method to adapt an LVCSR system to a specific domain based on stochastic segmentation. The proposed method is fully automatic. This means that the proposed method takes much less time than the ideal method. In addition, the proposed method yielded a comparable or even superior performance to the ideal method.

In conclusion, the proposed method allows us to adapt LVCSR to various domains in much less time.

7. ACKNOWLEDGMENTS

We thank the staff of *the University of the Air*.

8. REFERENCES

[1] K. Miyamoto, "Effective Master-Client Closed Caption Editing System for Wide Range Workforces," in *Proc. of HCI International 2005*.

[2] P. Fetter, "Detection and Transcription of OOV words," Tech. Rep. 231, Verbmobil, 8 1998.

[3] J. R. Bellegarda, "Statistical language model adaptation: review and perspectives," *Speech Communication*, vol. 42, pp. 93–108, 2004.

[4] D. Janiszek, R. De Mori, and F. Bechet, "Data Augmentation and Language Model Adaptation," in *Proc. of ICASSP 2001*, pp. 549–552.

[5] C. D. Manning and H. Schütze, *FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING*, THE MIT PRESS, 1999.

[6] M. Nagata, "A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm," in *Proc. of COLING 94*, pp. 201–207.

[7] S. Mori and D. Takuma, "Word N-gram Probability Estimation From A Japanese Raw Corpus," in *Proc. of ICSLP 2004*, pp. 201–207.

[8] H. Feng, K. Chen, X. Deng, and W. Zheng, "Accessor Variety Criteria for Chinese Word Extraction," *Computational Linguistics*, vol. 30, no. 1, pp. 75–93, 2004.

[9] M. Asahara and Y. Matsumoto, "Japanese Unknown Word Identification by Character-based Chunking," in *Proc. of COLING 2004*, pp. 459–465.

[10] G. Kurata, S. Mori, and M. Nishimura, "Large Vocabulary Continuous Speech Recognition with a Japanese Language Model from a Raw Corpus," 2005, 2005-SLP-57-19 (in Japanese).