UNSUPERVISED LEXICON ACQUISITION FROM SPEECH AND TEXT

Gakuto KURATA, Shinsuke MORI, Nobuyasu ITOH, Masafumi NISHIMURA

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

ABSTRACT

When introducing a Large Vocabulary Continuous Speech Recognition (LVCSR) system into a specific domain, it is preferable to add the necessary domain-specific words and their correct pronunciations selectively to the lexicon, especially in the areas where the LVCSR system should be updated frequently by adding new words. In this paper, we propose an unsupervised method of word acquisition in Japanese, where no spaces exist between words. In our method, by taking advantage of the speech of the target domain, we selected the domain-specific words among an enormous number of word candidates extracted from the raw corpora. The experiments showed that the acquired lexicon was of good quality and that it contributed to the performance of the LVCSR system for the target domain.

Index Terms— Large Vocabulary Continuous Speech Recognition, Stochastically segmented corpus, Lexicon acquisition

1. INTRODUCTION

Although a large general lexicon has been constructed, it can't cover all of the words in any domain[1]. In addition, many new words are appearing every day. Therefore, when introducing a Large Vocabulary Continuous Speech Recognition (LVCSR) system into a new domain, new words which are specific for that domain and which are not included in the general lexicon inevitably appear. Considering areas such as call centers and congress where the LVCSR system should be updated frequently by adding new words, we don't want to add many words into the lexicon each and every time, because the size of the lexicon of the system is limited, not infinite. Therefore, when introducing an LVCSR system into a specific domain, it is important to add the necessary domain-specific words selectively.

In Japanese, like some other Asian languages, no spaces exist between words. Identification of the domain-specific words from the raw corpora in specialized areas has been a difficult task[2] [1]. An automatic word segmenter also has problems at analyzing the domain-specific words because the automatic word segmenter itself is not trained with the domain-specific knowledge [2]. Therefore, even though raw corpora of the target domain are available, we can't extract the domain-specific words automatically from the raw corpora.

In this paper, we consider the situation of introducing an LVCSR system into a specific domain while adding the necessary domain-specific lexicons selectively. We assume that raw corpora and speech of the target domain are available. As is well known, many articles are computerized these days. In addition, speech data for the target domain is the very thing we are working with.

[1] In this paper, "lexicon" means a set of the pairs of a word and its pronunciation used in Large Vocabulary Continuous Speech Recognition.

[2] "Raw corpora" means a set of sentences that are not segmented into words.

In this paper, we propose an unsupervised method of word acquisition in Japanese. In an earlier paper, we proposed a method to add all of the probable character strings into the lexicon as domain-specific words [3]. Although these character strings contributed to improving the accuracy of the LVCSR system, most of them were just useless and meaningless character strings. In our proposed method, by taking advantage of the speech of the target domain, the domain-specific words are selected properly among the probable character strings extracted from the raw corpora. Corresponding pronunciations can be acquired simultaneously. The experiments showed that the acquired lexicon was of good quality and that the acquired lexicon contributed to the performance of the LVCSR system for the target domain.

2. PROPOSED METHOD

In this section, we describe our proposed method as shown in Figure 1. The key step in our method is acquisition of the domain-specific lexicon by integrating the speech and the raw corpora of the target domain. In this step, first, we extract an enormous number of word candidates from the raw corpora. Then we choose the appropriate domain-specific words through LVCSR over the speech of the target domain.In this way, we acquire a domain-specific lexicon that contains the appropriate domain-specific words and their pronunciations with high accuracy. After the lexicon acquisition, we build the LVCSR system for the target domain using this acquired lexicon.

On the assumption, we have large corpora of a general domain and a large lexicon based on these corpora. The general lexicon contains the general words and their pronunciations. In addition, we have a suitable Acoustic Model (AM). The details of these data will be described in Sec. 3. Please note that the general words are used in LM building and the AM, the general LM, and the general lexicon are used in recognition, but not depicted in Figure 1 to avoid confusion.

2.1. Lexicon Acquisition

We now describe how we acquire the domain-specific lexicon in detail. We also include an explanation using the domain-specific word "リン酸化" ("phosphation" in English) as an example. Please note that the numbers in Figure 1 correspond to the numbers of the following steps.

Step 1. Extraction of Word Candidates

We extracted the words from the raw corpora in order to collect the domain-specific words. As mentioned in Sec. 1, acquiring correct words from raw corpora is difficult. However, any words which are not in the lexicon for LVCSR will never be recognized. Therefore, we have to choose a method which achieves high recall. We used a traditional character-based approach to extract the probable character strings from the raw corpora. This approach is based on the frequencies of the character strings in the corpora [4, 5]. Because

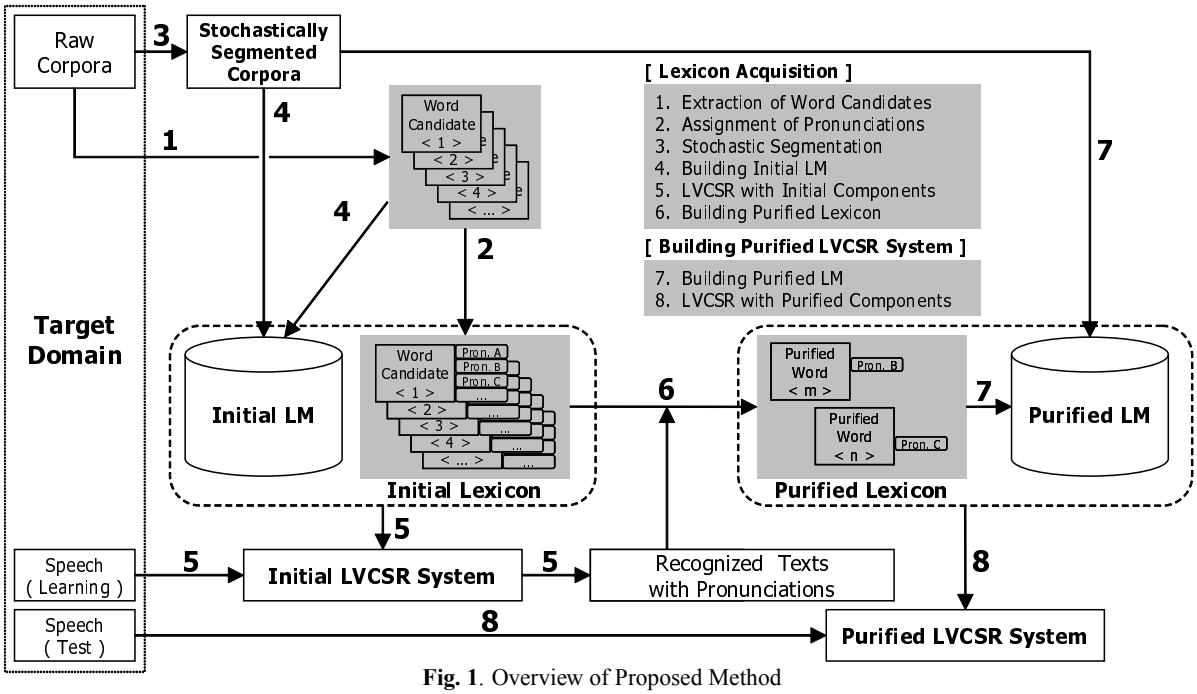

Fig. 1. Overview of Proposed Method

we focus on recall, many meaningless character strings tend to be selected in this step. We call them ***"Word Candidates"*** here.

Please assume that the domain-specific word "リン酸化" appears in the raw corpora and is extracted as a ***Word candidate***. Many other character strings are also extracted in this step.

Step 2. Assignment of Pronunciations

We need to assign pronunciations to the ***Word Candidates*** for LVCSR. In order to assign a pronunciation to a word that is not included in the general lexicon, the unknown word model is usually used, especially famous in the area of Text-To-Speech systems [6]. In the unknown word model, the pronunciation of a word is estimated based on a character n-gram model and a dictionary containing all possible pronunciations for each character. Unfortunately, the most probable pronunciation that the unknown word model selects is not always correct. Therefore, we assigned the ten most plausible pronunciations to each ***Word Candidate*** using the unknown word model. We call these pairs of ***Word Candidates*** and the assigned pronunciations the ***"Initial Lexicon"***.

Considering the word "リン酸化", the characters "リ" (/ri/) and "ン" (/n/) have only one pronunciation, but the characters "酸" (/sa n/ and /su/) and "化" (/ka/, /ke/, and /ba/) have multiple pronunciations, as written in parenthesis. As a result, the word "リン酸化" has 6 possible pronunciations as follows: /ri n sa n ka/, /ri n sa n ke/, /ri n sa n ba/, /ri n su ka/, /ri n su ke/, and /ri n su ba/. Only the pronunciation /ri n sa n ka/ is correct, but other pronunciations are also assigned here based on the spelling.

After Steps 1 and 2, we get an enormous number of ***Word Candidates*** and pronunciations. In the following steps, we will select the domain-specific words and their correct pronunciations from the ***Initial Lexicon*** through LVCSR over the speech of the target domain.

Step 3. Stochastic Segmentation

Steps 3 and 4 are the preparation for LVCSR in Step 5. Stochastic segmentation was proposed in [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 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"***.

In our experiments, the word boundary probabilities were defined as follows. First, the word boundary estimation accuracy α of an automatic word segmenter was calculated on a test corpus with word boundary information [2]. Then the raw corpus was segmented by the word segmenter. Finally p_i was set to be α for each i where the word segmenter put a word boundary and p_i was set to be $1 - \alpha$ for each i where it did not put a word boundary.

Step 4. Building Initial LM

A word n-gram model can be estimated from the list of words and the stochastically segmented corpora [7]. We built the word n-gram model for the target domain based on the ***Stochastically Segmented Corpora*** of the target domain, the word candidates, and the general words. We call the LM at this point the ***"Initial LM"***. We used a word tri-gram model throughout this paper.

Regarding the example, the probability $P_{LM}(\text{リン酸化}\,|w_h)$ is estimated here for each word history w_h.

Step 5. LVCSR with Initial Components

The LVCSR system for the target domain was constructed with the ***Initial Lexicon***, the ***Initial LM***, the general lexicon, the general LM, and the AM. We call this LVCSR system the ***"Initial LVCSR System"***. We split the speech of the target domain into two parts: a ***"Learning"*** part and a ***"Test"*** part. Then we had the ***Initial LVCSR System*** recognize the ***Learning*** part of the speech[3]. When a sufficient amount of raw corpora are available, the LVCSR system using stochastic segmentation achieves the best performance [3].

[3] We call this part of speech the ***"Learning"*** part, because we acquire the lexicon through LVCSR for this part. The ***"Test"*** part will be used for the evaluation.

Under the framework of LVCSR, the words are selected from the enormous number of ***Word Candidates*** in the ***Initial Lexicon*** when they satisfy the following two conditions.

- Their pronunciations appear in the ***Learning*** speech.
- Their contexts give them high LM probabilities.

Meaningless character strings and incorrect pronunciations don't satisfy these two conditions and are not selected.

Considering the example "リン酸化", if in the ***Learning*** speech, the phoneme sequence /ri n sa n ka/ appears in the contexts where the $P_{LM}(\text{リン酸化}\ |w_h)$ is high, the word "リン酸化" is selected from the enormous number of ***Word Candidates*** and its correct pronunciation /ri n sa n ka/ is selected.

Step 6. Building Purified Lexicon

By analyzing the recognized text, we picked up the words and their pronunciations that appeared in the recognized texts and belonged to the ***Initial Lexicon***.

The number of words and the number of pronunciations for each word decrease here compared with those in the ***Initial Lexicon*** because the ***Word candidates*** and their pronunciations are being verified through LVCSR. We call these selected ***Word candidates*** the ***"Purified Words"*** and the set of the pairs of ***Purified Words*** and their pronunciations the ***"Purified Lexicon"***. Analysis of this ***Purified Lexicon*** will be described in Sec. 5.1.

Looking at the example, the appropriate domain-specific word "リン酸化" is selected from the ***Word Candidates*** and the number of its pronunciations decreases from 6 to 1.

2.2. Building Purified LVCSR system

Now we have acquired the domain-specific lexicon, the ***Purified Lexicon***, through Steps 1 to 6. We will explain how to build an LVCSR system for the target domain using the ***Purified Lexicon***.

Step 7. Building Purified LM

We built the word n-gram model for the target domain based on the ***Purified Words***, the general words, and the same ***Stochastically Segmented Corpora***, as described above. We call this LM the ***"Purified LM"***.

Step 8. LVCSR with Purified Components

We constructed the LVCSR system for the target domain from the ***Purified Lexicon***, the ***Purified LM***, the general lexicon, the general LM, and the AM. We call this LVCSR system the ***"Purified LVCSR System"***.

3. EXPERIMENTAL SETUP

We briefly explain the AM, the general LM, and the general lexicon used in the experiment.

3.1. Acoustic Model

We used a spontaneous speech corpus of 83 hours to train the AM. Phones were represented as context-dependent, 3-state, left-to-right HMMs. The HMM states were clustered by a phonetic decision tree. The number of leaves was 2,728. Each state of the HMMs was modeled by 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. We built from this corpus a general LM and a general lexicon which were used in the experiment. The number of words in the general corpus was 24,442,503. The general lexicon contained 45,402 unique words and 53,225 pronunciations.

4. EXPERIMENTS

We conducted the experiments on a lecture 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.

We selected a lecture on biology. Table 1 shows the size of the raw corpora in relation to the lecture. These related corpora are mainly composed of the textbooks which are published by the *University of the Air*. The size of the raw corpora is approximately equivalent to that of one entire textbook. Table 1 also shows the size of the lecture speech, which was split into two parts, *Learning* and *Test*. The *Learning* part was used for word and pronunciation acquisition through the *Initial LVCSR System* and the *Test* part was used for the evaluation.

Table 1. Statistics of the Lecture

Raw Corpora	Speech [min.]	
[# characters]	Learning	Test
73,437	12.3	6.2

We built the *Purified Lexicon* for the target lecture according to the proposed method described in Sec. 2 and investigated it. Then in order to confirm that the acquired lexicon contributes to LVCSR, we built the *Purified LVCSR System* and used it to recognize the *Test* speech.

5. EVALUATION AND DISCUSSION

We explain the results of the experiments and discuss them.

5.1. Purified Lexicon

We examined the *Purified Lexicon* and calculated its accuracy. We regarded the pair of a word and its pronunciation as correct when the word is an appropriate domain-specific word and its pronunciation is correct. Regarding the compound words, we judged them according to their dependency structures [1]. Table 2 shows the accuracy of the *Purified Lexicon*. We calculated three accuracy metrics. The first column shows the accuracy for the words which appeared more than once in the recognized texts; the second column shows the accuracy for the words which appeared only once; and the rightmost column shows the total accuracy for all of the words in the *Purified Lexicon*.

Table 2. Accuracy of *Purified Lexicon* [%]

More than Once	Once	Total
97.2	68.9	79.5

The accuracy of "More than Once" is much higher than that of "Once". Appearing in the recognized texts more than once means that the word was spoken with the corresponding pronunciation, appeared in the contexts with high LM probability and was used in the lecture on the target domain more than once. In contrast, appearing only once can be an accidental insertion or substitution error. Therefore, this difference is reasonable and using only the words appearing multiple times seems to be a good method when a sufficient amount of *Learning* speech is available.

Table 3 shows some examples from the *Purified Lexicon*.

Overall the *Purified Lexicon* is of good quality and we can expect it to contribute to the LVCSR system for the target domain.

Table 3. Examples of *Purified Lexicon*

Frequency	Word (English Translation)	Pronunciation
27	受容体 (receptor)	ju yo o ta i
13	リン酸化 (phosphation)	ri n sa n ka
12	サブユニット (subunit)	sa bu yu ni tto
2	単量体 (monomer)	ta n ryo o ta i
2	残基 (residue)	za n ki

⋆ "Frequency" means the number of times in the recognized texts.

5.2. Purified LVCSR System

First, we explain the criterion for evaluation. To measure the recognition accuracy, we used the Character Error Ratio (CER). The reason is that ambiguity exists in word segmentation in Japanese. 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 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, so the Word Error Rate (WER) fluctuates accordingly. Therefore, the CER is a suitable criterion in Japanese. For reference, we estimated the WER based on the CER and the average number of characters $\bar{n}$ per word. We named this criterion "*e*WER" and this was defined as follows : $e\text{WER} = (1 - (1 - \text{CER})^{\bar{n}})$.

We compared the *Purified* and the *Initial LVCSR Systems* for their recognition accuracies. For reference, we built an LVCSR system with the general lexicon, the general LM, and the AM and had it recognize the speech of the target domain. We call this LVCSR system the "General LVCSR system". The results and the other statistics are shown in Table 4.

Table 4. Comparison of *Initial* and *Purified LVCSR Systems*

LVCSR System	Lexicon # Words <# Pronunciations>	CER (*e*WER) [%] < OOV rate [%] >	
		Learning	Test
General	— < — >	27.0 (53.0) < 5.88 >	26.1 (51.6) < 6.15 >
Initial	3,999 < 26,169 >	12.3 (27.0) < 2.31 >	**9.9 (22.0)** < 1.71 >
Purified	326 < 326 >	11.7 (25.8) < 2.31 >	**9.7 (21.7)** < 1.90 >

The second column of Table 4 shows the added lexicon for each LVCSR system. The numbers of words and pronunciations in the *Purified Lexicon* were much smaller than that in the *Initial Lexicon* because only the lexicon verified with the *Learning* speech was included in the *Purified Lexicon*.

The third column shows the CER, the *e*WER, and the Out-Of-Vocabulary (OOV) rate for the *Learning* speech. The performance of the *Purified LVCSR System* was superior to that of the *Initial LVCSR System*. Although the *Learning* speech was the closed data for the *Purified LVCSR System*, we improved the performance automatically. The reason for the improved performance is that the probabilities are properly distributed to the domain-specific words in the *Purified LM*, compared with the *Initial LM* in which the probabilities were widely distributed to the many meaningless character strings. In other words, when the speech of the target domain is fixed, we can improve the performance of the LVCSR system by using all of the speech data as the *Learning* speech in the framework of our proposed method.

The rightmost column shows the results for the *Test* speech, which was the open data for all LVCSR systems. Comparing the CERs, the performance of the *Purified LVCSR System* was comparable to that of the *Initial LVCSR System* even though the number of the words in the *Purified Lexicon* decreased from 3,999 to 326 and the number of pronunciations decreased from 26,169 to 326. This indicates that the *Purified Lexicon* contains appropriate domain-specific words and coincides with the knowledge we described in Sec. 5.1. This result shows that by leveraging the speech of the target domain, we can build an LVCSR system for the target domain while adding a smaller lexicon.

Considering the OOV rate, the OOV rate of the *Test* speech for the *Purified LVCSR System* increased. The reason for this is that only the recognized words for the *Learning* speech with the *Initial LVCSR System* were included in the *Purified Lexicon*. In this case, the *Learning* speech didn't cover all of the domain-specific words in the raw corpora. In order to get the best out of the proposed method, the *Learning* speech needs to cover as many domain-specific words as possible in the raw corpora.

6. CONCLUSION

In this paper, we proposed an unsupervised method of word acquisition in Japanese. In our method, by taking advantage of the speech of the target domain, we selected the domain-specific words among an enormous number of word candidates extracted from the raw corpora. We confirmed that the acquired lexicon was of good quality and that the acquired lexicon contributed to the performance of the LVCSR system for the target domain.

Though the size of the speech data and the raw corpora in the experiments we conducted were not so large, the results were promising. In our method, the coverage of the speech data for the raw corpora is critical, just as the coverage of the corpus for the speech has been important for LVCSR. In addition, it should be beneficial to use only the words that appear multiple times in the *Initial* recognition when the *Learning* speech data is sufficient. Therefore, when larger raw corpora and more speech are available, our proposed method has the potential to work even more effectively.

7. ACKNOWLEDGEMENTS

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

8. REFERENCES

[1] M. Asahara *et al.*, "Japanese Unknown Word Identification by Character-based Chunking," in *Proc. COLING*, 2004, pp. 459–465.

[2] M. Nagata, "A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm," in *Proc. COLING*, 1994, pp. 201–207.

[3] G. Kurata *et al.*, "Unsupervised Adaptation of a Stochastic Language Model Using a Japanese Raw Corpus," in *Proc. ICASSP*, 2006, vol. 1, pp. 1037–1040.

[4] H. Feng *et al.*, "Accessor Variety Criteria for Chinese Word Extraction," *Computational Linguistics*, vol. 30, no. 1, pp. 75–93, 2004.

[5] G. Kurata *et al.*, "Large Vocabulary Continuous Speech Recognition with a Japanese Language Model from a Raw Corpus," 2005, 2005-SLP-57-19 (in Japanese).

[6] T. Nagano *et al.*, "A Stochastic Approach to Phoneme and Accent Estimation," in *Proc. Interspeech*, 2005.

[7] S. Mori *et al.*, "Word N-gram Probability Estimation From A Japanese Raw Corpus," in *Proc. Interspeech*, 2004.