# Structured language model for class identification of out-of-vocabulary words arising from multiple word-classes

*Shigehiko Onishi, Hirofumi Yamamoto, Yoshinori Sagisaka*

ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho, Soraku-gun, Kyoto 619-0288 Japan
{sonishi, yama, sagisaka}@slt.atr.co.jp

## Abstract

A structured language model (STLM) is proposed to cope with out-of-vocabulary (OOV) words coming from multiple word-classes. The STLM aims at independently modeling the classes without interference and identifying the class of words arising from multiple word-classes. The STLM consists of the conventional word-class N-gram and the sets of the independent-trained class-specific sub-word N-grams. We made an experimental language model by using STLM for the two similar proper-noun classes and performed the speech recognition experiments. The results show that any OOV word of the one class is never misrecognized as that of the other class. The results show that the STLM could integrate the multiple different statistical language models with no interference.

## 1. Introduction

It is well known that out-of-vocabulary (OOV) words often cause serious damage to the speech recognition results. OOV words and their neighbors are misrecognized to one or more different words with similar acoustic profiles. The OOV problem has been approached by the simple expansion of vocabulary, and this led to the development of large-vocabulary continuous speech recognition systems. However, it would be difficult to cover OOV words such as proper nouns whose numbers are huge.

OOV words have been detected and rejected to reduce neighboring word recognition errors. Recently, word-class specific OOV modeling has been pursued not for rejection but for identifying word-classes like city name [1], number [2] and Japanese family/personal name [3]. Even if the phonetic transcription of OOV words is mistaken, the correct word-class information is useful for posterior semantic processing such as the correct handling of human names in automatic speech translation.

In word-class identification, all previous works targeted only a single word-class. Those are insufficient for general use. Even in one easy sentence, OOV words could arise from the multiple word-classes especially for proper nouns. For example, a sentence “Mr. OOV(surname) lives in OOV(city-name)” has the two OOV words one from the class surname and the other from the class city-name. If these classes were misidentified, the sentence would have a strange meaning. It is important to handle OOV words without mutual interference for identifying those word-classes.

For this purpose, in this paper, we propose a structured language model (STLM) to cope with OOV words that come from multiple word-classes. For a single class, we have already proposed a hierarchical language model [3] that has been proved to contribute for the identification of OOV words without damaging recognition of registered words. To cover all plausible OOV words, this model should be expanded to multiple word-classes. We propose a new word-structure model aiming at general modeling of the word-classes by using independent class-specific data. Speech recognition results are shown on the effectiveness of this modeling of OOV words from two word-classes without any interference between these OOV models.

## 2. Structured language model (STLM)

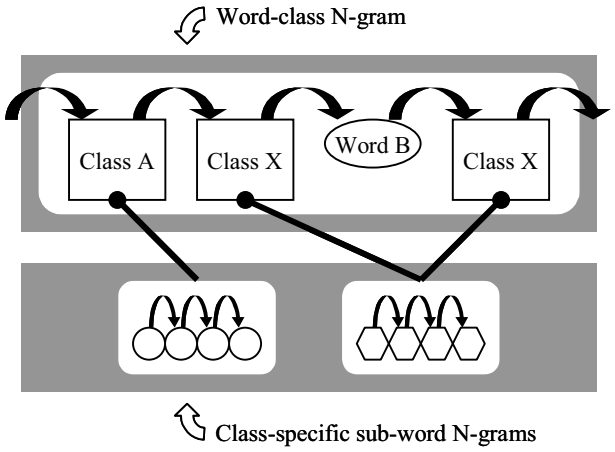


*Figure 1: The schematic of structured language model (STLM). This model consists of two layers and each of those provides different statistical constraints.*

A structured language model (STLM) has two layers as shown in Figure 1. The upper layer is the conventional word-class N-gram [4] that represents the word- and class-level statistical constraints in some domain. The lower layer consists of the sets of the sub-word N-grams. Each of them corresponds to a word-class of the upper layer and gives the class-specific statistical phonetic constraints of all words of that class. The STLM could integrate these models to one statistical language model independently and aims at providing independent statistical constraints to all words.

Each sub-word model is independently trained by using rather general corpus of the corresponding class and

represents each word as a class-specific sub-word sequence. If only the sub-word model could generate the infinite number of sub-word sequences, OOV words of the class could be covered, and all words in the class could be given their emission probability with the two kinds of estimation independently as (1).

$$P(c_N \mid c_1 \cdots c_{N-1}) P(R \mid c_N) \qquad (1)$$

In (1), an OOV word in a class $c_N$ is represented by a sub-word sequence $R$. The estimation $P(c_N \mid c_1 \cdots c_{N-1})$ and $P(R \mid c_N)$ are given by the upper and the lower layer model, respectively.

For general modeling of the sub-word models, a new design method has been needed for covering all sub-word sequences and independently giving the class-specific estimation $P(R \mid c_N)$. For this purpose, we propose the word-structure model as described in the following sections.

### 2.1. Word-structure model for the class-specific sub-word N-gram

To get a good estimation of $P(R \mid c_N)$, it is important to design the class-specific sub-word N-gram to catch the features of corresponding word-class $c_N$. This is true especially for STLM; since the STLM would consist of many sub-word models, the expressive model that can reflect slight differences of the class features is desired. Thus, in this paper, we propose a model based on the assumption that there are specific phonetic constraints at the beginning and at the ending of the sub-word sequence. This model is named word-structure model and can be formed like a probabilistic finite-state automaton (PFSA) that has infinite coverage of sub-word sequences. With the word-structure model, the estimation precision of $P(R \mid c_N)$ can be easily enhanced. Regarding a connected sub-word as a single one, the number of sub-word units is increased, and it would provide the enhancement of the estimation precision [5]. From this property, a word-structure model can be adaptively improved to reflect the difference of the class features.

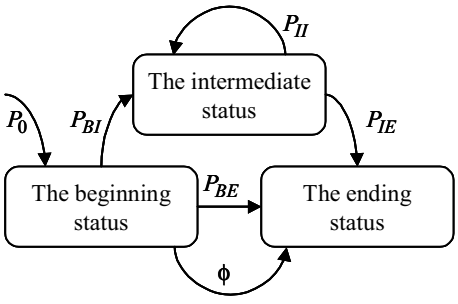


*Figure 2: A word-structure model. This model simulates the sub-word sequence of the word. Each of the arcs corresponds to the sub-word bi-gram. (the arc labeled $\phi$ means null transition)*

A word-structure model is derived as follows. Let the sub-word sequence $R$ equals $a_B a_2 \cdots a_i \cdots a_{m-1} a_E$, where $a_B$ and $a_E$ are the beginning and the ending sub-word respectively, and $a_2, \cdots, a_i, \cdots, a_{m-1}$ are the sub-word suffixed with its position in $R$. Considering position dependency of the sub-words, the bi-gram model of $R$ becomes as (2).

$$P(R \mid c_N) \approx P_0(a_B \mid start, c_N) P_1(a_2 \mid a_B, c_N) P_2(a_3 \mid a_2, c_N) \cdots \cdots P_{N-2}(a_{N-1} \mid a_{N-2}, c_N) P_{N-1}(a_E \mid a_{N-1}, c_N) \qquad (2)$$

In (2), notation $P_n(y \mid x)$ means the bi-gram from x to y where x and y are in n'th and (n+1)'th position, respectively. Since this model needs too many bi-gram parameters for training on reasonable corpora, the simpler model would be needed. Approximating that only the bi-grams that include the beginning and/or the ending sub-word have the strict value and the remaining bi-grams are calculated by neglecting the position of the sub-word, then, equation (2) becomes (3).

$$P(R \mid c_N) \approx P_0(a_B \mid start, c_N) P_{BI}(a_2 \mid a_B, c_N) P_{IE}(a_E \mid a_{N-1}, c_N) \times \prod P_{II}(a_{i+1} \mid a_i) \qquad (3)$$

This is the formula of word-structure model. Where $a_2$, $a_i$, $a_{i+1}$ and $a_{N-1}$ are the intermediate sub-words those appear at neither the beginning nor the ending of $R$. The notation $P_{BI}(a_2 \mid a_B)$ means the from-beginning-to-intermediate bi-gram from the beginning sub-word $a_B$ to the intermediate sub-word $a_2$. Similarly, the notation $P_{IE}(.\mid.)$ and $P_{II}(.\mid.)$ mean from-intermediate-to-ending and from-intermediate-to-intermediate bi-grams, respectively. Note that, equation (3) allows any number of productions of from-intermediate-to-intermediate bi-grams. This gives infinite coverage of sub-word sequences. In special case that the sequence $R$ has only one or two sub-words, equation (3) is denoted as follows:

$R$ has two sub-words:

$$P(R \mid c_N) \approx P_0(a_B \mid start, c_N) P_{BE}(a_E \mid a_B, c_N)$$

$R$ has only one sub-word:

$$P(R \mid c_N) \approx P_0(a_{BE} \mid start, c_N)$$

Where, the notation $P_{BE}(.\mid.)$ means the from-beginning-to-ending bi-gram and the $a_{BE}$ is the beginning-ending sub-word.

### 2.2. Implementation of word-structure model

As equation (3), the formula of the word-structure model is almost the same as that of a conventional bi-gram model. Also, the implementation of this is also the same except for the special labeling rules and the restrictions to simulate the PFSA shown in Figure 2.

After converting each training word to a sub-word sequence (the sub-word could be chosen as phoneme, syllable, or other phonetic unit), the two kinds of labels are attached. The first is for identifying the class-name affixed to all sub-words. And the second is for specifying the edge of the word affixed to the beginning/ending/beginning-ending sub-word of the sequence. For example, a Japanese word "akai" has three syllables and converted by the operation above to "BCCa

CCka ECCi", where CC is the class-identifying label, B and E are the beginning and the ending label, respectively.

Using the sub-word corpus mentioned above, bi-gram parameters could be easily obtained by conventional training scheme. Then, the restrictions for simulating the PFSA are applied. That is, the following bi-grams are eliminated: from-intermediate-to-beginning bi-grams, from-ending-to-intermediate bi-grams, and from-ending-to-beginning bi-grams.

### 2.3. Experimental word-structure models

Experimental word-structure models were trained on a corpus of the class Japanese family/personal names (598k words, 39k vocabulary) with various numbers of sub-word units. Those units were syllables and their connections. The perplexity of each model was measured by using a test set of the same class (2k syllables) and the results are shown in Figure 3. Increasing the number of units monotonically decreases the perplexity. This means that the model becomes precise by increasing the sub-word units.

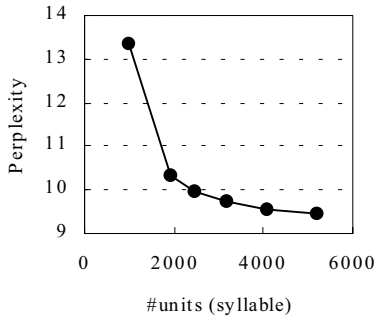


*Figure 3: Test set perplexity of word-edge models made for the word-class Japanese family/personal names with various numbers of sub-word (syllable) units.*

## 3. Speech recognition experiments

To confirm the effectiveness of the modeling for the OOV words from multiple word-classes, speech recognition experiments were performed using data including OOV words from two different word-classes. The experiments were carried out by comparison among three language models: the baseline language model, the STLM modeling one class, and the STLM modeling two classes.

### 3.1. The choice of the target classes

We chose two target classes: Japanese family/personal names (JF/PN) and Japanese place names (JZP). The class JF/PN is modeled in our previous works [3] and the class JZP is selected since place names would be widely used for our recognition tasks like travel arrangement and car navigation. These two classes are lexically similar. They have many common sub-word (phoneme) sequences. For example, a phoneme sequence "m/o/r/i" is exists in both classes, as the surname in JF/PN and as the town's name in JZP.

The volume of the experimental corpora of those classes is shown in Table 1. The JF/PN corpus was collected from the database on the market [6] and the JZP corpus was extracted from the address data listed on the public postal-code database [7]. Each of the databases was processed by eliminating common suffixes, for example, "-ken" (prefecture), "-shi" (city), etc.

| Corpus size | JF/PN | JZP |
|---|---|---|
| Words | 598k | 405k |
| Vocabulary | 39k | 51k |

*Table 1: The experimental corpora of the class JF/PN and JZP*

### 3.2. Procedure of experiments

For the comparative experiments, the following three language models were made:

1. A conventional word-class language model (LM_Base)

   The word-class language model [4] trained on the travel arrangement corpus (1161k words, 13k vocabulary) [8]. This model gives the baseline of the total recognition rates and is also the basis for building the upper layer of STLM of following models.

2. A STLM modeling one word-class (STLM-1)

   This is a STLM modeling the class JF/PN. The upper layer of this model is made by eliminating all JF/PN-word entries from LM_Base, while the lower layer is trained on JF/PN corpus shown in Table 1 by using word-structure model. Note that the lower model has 5212 units (syllables and connected syllables) and perplexity is 9.5. This model gives the baseline of the recognition rates only for all JF/PN words.

3. A STLM modeling two word-classes (STLM-2)

   This is a STLM modeling both the class JF/PN and JZP. The constructing operation is similar to above. The upper is made by eliminating all JZP words from STLM-1, and the lower is trained on JZP corpus. The lower model has 2856 units (syllables and connected syllables) and perplexity is 10.5. This model is used to verify whether two word-classes interfered with each other.

Note that all JF/PN words are represented as sub-word sequences in STLM-1 and STLM-2. Also, all JZP words are given as sequences only in STLM-2. The following criteria were used for correct recognition of these sequences: correct class information, correct sub-word sequence, and correct locations (by DP-matching).

For speech recognition, ATRSPREC [9] was used. The test set has 1289 utterances (16k words) selected from the same domain as that of the training set for LM_Base. It has 171 JF/PN words and 46 JZP words.

## 4. Results and Discussions

Table 2 shows the counts of the JF/PN and JZP words that were substituted by the others when using STLM-2. The

counts are classified by the class of the substitute words. It is obvious that no JF/PN word was ever identified as JZP word. Also no JZP word was recognized as JF/PN word. From this result, it is verified that there is no interference in word-class identification between two classes JF/PN and JZP.

| | | Class of correct words | |
|---|---|---|---|
| | | JF/PN | JZP |
| Class of substitute words | JF/PN | 17 | **0** |
| | JZP | **0** | 10 |
| | Others | 13 | 22 |

*Table 2: Counts of the substituted words of the class JF/PN and JZP. JF/PN words and JZP words never interfered with each other.*

Table 3 shows the recognition rate in terms of word accuracy for whole words. From this result, no degradation was found by using STLM. This results supports that STLM could apply to multiple word-classes without damaging recognition of other registered words.

| Language model | Word accuracy |
|---|---|
| LM_Base | 86.6 |
| STLM-1 | 87.0 |
| STLM-2 | 87.7 |

*Table 3: Word accuracy for whole words. STLM never degrades the performance.*

Table 4 shows the precision/recall rate of JF/PN and JZP words. The following results were found:

- The recall rates of JF/PN words were 70% for STLM-1 and 73% for STLM-2. This increase is due to the fact that some JF/PN words were included in the words that are substituted for by the JZP words. Since all JZP words were converted to sub-word sequences in STLM-2, the JF/PN words masked by JZP words would appear in the results.
- The precision rates of JF/PN words were 83% for STLM-1 and 81% for STLM-2. The reason of this degradation would be same as above. Since some words (not JF/PN words) masked by the JZP sub-words are originally mistaken, those would be substituted for by the JF/PN sub-words.
- For STLM-2, the precision rate of JZP words was rather lower compared with that of JF/PN words. We examined all errors caused by JZP words, and found that they occur in the region that is originally damaged. No new error was found.

From these results, we found the effectiveness of STLM for modeling OOV words from multiple word-classes without interference and degradation of recognition rate.

| Language model | P/R for JF/PN | P/R for JZP |
|---|---|---|
| STLM-1 | 83/70 | -/- |
| STLM-2 | 81/73 | 49/65 |

*Table 4: Precision/Recall rate of JF/PN and JZP words. The recall rate of JF/PN words was enhanced from 70% to 73% by modeling JZP words.*

## 5. Conclusion

We propose a structured language model (STLM) to cope with OOV words come from multiple word-classes. The STLM aims at independently modeling the word-classes without interference and identifying the class of words arising from multiple classes. The STLM has a task-dependent word-class N-gram as the upper layer and sets of independently trained class-specific sub-word N-grams as the lower layer. We made an experimental language model by using STLM for the two similar classes, Japanese family/personal names and Japanese place names, and performed speech recognition experiments. The results show that any OOV word of the one class is never misrecognized as that of the other class. The results show that STLM could integrate the independent statistical models without interference. We plan to extend this model to many kinds of proper nouns, and aim at constructing general language model for task-independent automatic speech recognition systems.

## 6. References


[1] F. Gallwitz, E. Nöth and H. Niemann, “A class based approach for recognition of out-of-vocabulary words”, Proc. ICSLP 1996, pp.228-231.

[2] L. Galescu and J. Allen, “Hierarchical statistical language models: experiments on in-domain adaptation”, ICSLP 2000, pp.186-189.

[3] K. Tanigaki, H. Yamamoto and Y. Sagisaka, “A hierarchical language model incorporating class-dependent word models for OOV words recognition”, Proc. ICSLP 2000, pp.123-126.

[4] H. Yamamoto and Y. Sagisaka, “Multi-class composite N-gram based on connection direction”, Proc. ICASSP 1999, pp.533-536.

[5] H. Masataki and Y. Sagisaka, “Variable-order N-gram generation by word-class splitting and consecutive word grouping”, Proc. ICASSP 1996, vol.1, pp.188-191.

[6] Nichigai Associates Incorporated, “EB reading and writing dictionary of 300 thousand peoples”, ISBN 4-8169-7020-7, 1993.

[7] http://www.postal.mpt.go.jp/newnumber/lzh/s/ken_all.lzh, data download services of The Ministry of Posts and Telecommunications of Japan.

[8] A. Nakamura et al., “Japanese speech databases for robust speech recognition”, Proc. ICSLP 1996, pp.2199-2202.

[9] T. Shimizu and Y. Sagisaka, “Fast word-graph generation for spontaneous conversational speech translation”, Proc. ICASSP 1997, pp.95-98.