# Multi-Class Composite N-gram Language Model for Spoken Language Processing Using Multiple Word Clusters

**Hirofumi Yamamoto**
ATR SLT
2-2-2 Hikaridai Seika-cho
Soraku-gun, Kyoto-fu, Japan
yama@slt.atr.co.jp

**Shuntaro Isogai**
Waseda University
3-4-1 Okubo, Shinjuku-ku
Tokyo-to, Japan
isogai@shirai.info.waseda.ac.jp

**Yoshinori Sagisaka**
GITI / ATR SLT
1-3-10 Nishi-Waseda
Shinjuku-ku, Tokyo-to, Japan
sagisaka@slt.atr.co.jp

## Abstract

In this paper, a new language model, the Multi-Class Composite N-gram, is proposed to avoid a data sparseness problem for spoken language in that it is difficult to collect training data. The Multi-Class Composite N-gram maintains an accurate word prediction capability and reliability for sparse data with a compact model size based on multiple word clusters, called Multi-Classes. In the Multi-Class, the statistical connectivity at each position of the N-grams is regarded as word attributes, and one word cluster each is created to represent the positional attributes. Furthermore, by introducing higher order word N-grams through the grouping of frequent word successions, Multi-Class N-grams are extended to Multi-Class Composite N-grams. In experiments, the Multi-Class Composite N-grams result in 9.5% lower perplexity and a 16% lower word error rate in speech recognition with a 40% smaller parameter size than conventional word 3-grams.

## 1 Introduction

Word N-grams have been widely used as a statistical language model for language processing. Word N-grams are models that give the transition probability of the next word from the previous $N-1$ word sequence based on a statistical analysis of the huge text corpus. Though word N-grams are more effective and flexible than rule-based grammatical constraints in many cases, their performance strongly depends on the size of training data, since they are statistical models.

In word N-grams, the accuracy of the word prediction capability will increase according to the number of the order N, but also the number of word transition combinations will exponentially increase. Moreover, the size of training data for reliable transition probability values will also dramatically increase. This is a critical problem for spoken language in that it is difficult to collect training data sufficient enough for a reliable model. As a solution to this problem, class N-grams are proposed.

In class N-grams, multiple words are mapped to one word class, and the transition probabilities from word to word are approximated to the probabilities from word class to word class. The performance and model size of class N-grams strongly depend on the definition of word classes. In fact, the performance of class N-grams based on the part-of-speech (POS) word class is usually quite a bit lower than that of word N-grams. Based on this fact, effective word class definitions are required for high performance in class N-grams.

In this paper, the Multi-Class assignment is proposed for effective word class definitions. The word class is used to represent word connectivity, i.e. which words will appear in a neighboring position with what probability. In Multi-Class assignment, the word connectivity in each position of the N-grams is regarded as a different attribute, and multiple classes corresponding to each attribute are assigned to each word. For

the word clustering of each Multi-Class for each word, a method is used in which word classes are formed automatically and statistically from a corpus, not using a priori knowledge as POS information. Furthermore, by introducing higher order word N-grams through the grouping of frequent word successions, Multi-Class N-grams are extended to Multi-Class Composite N-grams.

## 2 N-gram Language Models Based on Multiple Word Classes

### 2.1 Class N-grams

Word N-grams are models that statistically give the transition probability of the next word from the previous $N-1$ word sequence. This transition probability is given in the next formula.

$$p(w_i|w_{i-N+1}, ..., w_{i-2}, w_{i-1}) \quad (1)$$

In word N-grams, accurate word prediction can be expected, since a word dependent, unique connectivity from word to word can be represented. On the other hand, the number of estimated parameters, i.e., the number of combinations of word transitions, is $V^N$ in vocabulary $V$. As $V^N$ will exponentially increase according to $N$, reliable estimations of each word transition probability are difficult under a large $N$.

Class N-grams are proposed to resolve the problem that a huge number of parameters is required in word N-grams. In class N-grams, the transition probability of the next word from the previous $N-1$ word sequence is given in the next formula.

$$p(c_i|c_{i-N+1}, ..., c_{i-2}, c_{i-1})p(w_i|c_i) \quad (2)$$

Where, $c_i$ represents the word class to which the word $w_i$ belongs.

In class N-grams with $C$ classes, the number of estimated parameters is decreased from $V^N$ to $C^N$. However, accuracy of the word prediction capability will be lower than that of word N-grams with a sufficient size of training data, since the representation capability of the word dependent, unique connectivity attribute will be lost for the approximation base word class.

### 2.2 Problems in the Definition of Word Classes

In class N-grams, word classes are used to represent the connectivity between words. In the conventional word class definition, word connectivity for which words follow and that for which word precedes are treated as the same neighboring characteristics without distinction. Therefore, only the words that have the same word connectivity for the following words and the preceding word belong to the same word class, and this word class definition cannot represent the word connectivity attribute efficiently. Take "a" and "an" as an example. Both are classified by POS as an Indefinite Article, and are assigned to the same word class. In this case, information about the difference with the following word connectivity will be lost. On the other hand, a different class assignment for both words will cause the information about the community in the preceding word connectivity to be lost. This directional distinction is quite crucial for languages with reflection such as French and Japanese.

### 2.3 Multi-Class and Multi-Class N-grams

As in the previous example of "a" and "an", following and preceding word connectivity are not always the same. Let's consider the case of different connectivity for the words that precede and follow. Multiple word classes are assigned to each word to represent the following and preceding word connectivity. As the connectivity of the word preceding "a" and "an" is the same, it is efficient to assign them to the same word class to represent the preceding word connectivity, if assigning different word classes to represent the following word connectivity at the same time. To apply these word class definitions to formula (2), the next formula is given.

$$p(c_i^t|c_{i-N+1}^{fN-1}, ..., c_{i-2}^{f2}, c_{i-1}^{f1})p(w_i|c_i^t) \quad (3)$$

In the above formula, $c_i^t$ represents the word class in the target position to which the word $w_i$ belongs, and $c_i^{fN}$ represents the word class in the N-th position in a conditional word sequence.

We call this multiple word class definition, a Multi-Class. Similarly, we call class N-grams based on the Multi-Class, Multi-Class N-grams (Yamamoto and Sagisaka, 1999).

## 3 Automatic Extraction of Word Clusters

### 3.1 Word Clustering for Multi-Class 2-grams

For word clustering in class N-grams, POS information is sometimes used. Though POS information can be used for words that do not appear in the corpus, this is not always an optimal word classification for N-grams. The POS information does not accurately represent the statistical word connectivity characteristics. Better word-clustering is to be considered based on word connectivity by the reflection neighboring characteristics in the corpus. In this paper, vectors are used to represent word neighboring characteristics. The elements of the vectors are forward or backward word 2-gram probabilities to the clustering target word after being smoothed. And we consider that word pairs that have a small distance between vectors also have similar word neighboring characteristics (Brown et al., 1992) (Bai et al., 1998). In this method, the same vector is assigned to words that do not appear in the corpus, and the same word cluster will be assigned to these words. To avoid excessively rough clustering over different POS, we cluster the words under the condition that only words with the same POS can belong to the same cluster. Parts-of-speech that have the same connectivity in each Multi-Class are merged. For example, if different parts-of-speeche are assigned to "a" and "an", these parts-of-speeche are regarded as the same for the preceding word cluster. Word clustering is thus performed in the following manner.

1. Assign one unique class per word.s.

2. Assign a vector to each class or to each word $X$. This represents the word connectivity attribute.

$$v^t(x) = [p^t(w_1|x), p^t(w_2|x), ..., p^t(w_N|x)] \quad (4)$$

$$v^f(x) = [p^f(w_1|x), p^f(w_2|x), ..., p^f(w_N|x)] \quad (5)$$

Where, $v^t(x)$ represents the preceding word connectivity, $v^f(x)$ represents the following word connectivity, and $p^t$ is the value of the probability of the succeeding class-word 2-gram or word 2-gram, while $p^f$ is the same for the preceding one.

3. Merge the two classes. We choose classes whose dispersion weighted with the 1-gram probability results in the lowest rise, and merge these two classes:

$$U_{new} = \sum_w (p(w)D(v(c_{new}(w)), v(w))) \quad (6)$$

$$U_{old} = \sum_w (p(w)D(v(c_{old}(w)), v(w))) \quad (7)$$

where we merge the classes whose merge cost $U_{new} - U_{old}$ is the lowest. $D(v_c, v_w)$ represents the square of the Euclidean distance between vector $v_c$ and $v_w$, $c_{old}$ represents the classes before merging, and $c_{new}$ represents the classes after merging.

4. Repeat step 2 until the number of classes is reduced to the desired number.

### 3.2 Word Clustering for Multi-Class 3-grams

To apply the multiple clustering for 2-grams to 3-grams, the clustering target in the conditional part is extended to a word pair from the single word in 2-grams. Number of clustering targets in the preceding class increases to $V^2$ from $V$ in 2-grams, and the length of the vector in the succeeding class also increase to $V^2$. Therefore, efficient word clustering is needed to keep the reliability of 3-grams after the clustering and a reasonable calculation cost.

To avoid losing the reliability caused by the data sparseness of the word pair in the history of 3-grams, approximation is employed using distance-2 2-grams. The authority of this approximation is based on a report that the association of word 2-grams and distance-2 2-grams based on the maximum entropy method gives a good approximation of word 3-grams (Zhang et al., 1999). The vector for clustering is given in the next equation.

$$v^{f2}(x) = [p^{f2}(w_1|x), p^{f2}(w_2|x), ..., p^{f2}(w_N|x)] \quad (8)$$

Where, $p^{f2}(y|x)$ represents the distance-2 2-gram value from word $x$ to word $y$. And the POS constraints for clustering are the same as in the clustering for preceding words.

## 4 Multi-Class Composite N-grams

### 4.1 Multi-Class Composite 2-grams Introducing Variable Length Word Sequences

Let's consider the condition such that only word sequence $(A, B, C)$ has sufficient frequency in sequence $(X, A, B, C, D)$. In this case, the value of word 2-gram $p(B|A)$ can be used as a reliable value for the estimation of word $B$, as the frequency of sequence $(A, B)$ is sufficient. The value of word 3-gram $p(C|A, B)$ can be used for the estimation of word $C$ for the same reason. For the estimation of words $A$ and $D$, it is reasonable to use the value of the class 2-gram, since the value of the word N-gram is unreliable (note that the frequency of word sequences $(X, A)$ and $(C, D)$ is insufficient). Based on this idea, the transition probability of word sequence $(A, B, C, D)$ from word $X$ is given in the next equation in the Multi-Class 2-gram.

$$\begin{aligned} P &= p(c^t(A)|c^f(X))p(A|c^t(A))) \\ &\times p(B|A) \\ &\times p(C|A, B) \\ &\times p(c^t(D)|c^f(C))p(D|c^t(D)) \end{aligned} \quad (9)$$

When word succession $A+B+C$ is introduced as a variable length word sequence $(A, B, C)$, equation (9) can be changed exactly to the next equation (Deligne and Bimbot, 1995) (Masataki et al., 1996).

$$\begin{aligned} P &= p(c^t(A)|c^f(X))p(A+B+C|c^t(A)) \\ &\times p(c^t(D)|c^f(C))p(D|c^t(D)) \end{aligned} \quad (10)$$

Here, we find the following properties. The preceding word connectivity of word succession $A+B+C$ is the same as the connectivity of word $A$, the first word of $A+B+C$. The following connectivity is the same as the last word $C$. In these assignments, no new cluster is required. But conventional class N-grams require a new cluster for the new word succession.

$$c^t(A+B+C) = c^t(A) \quad (11)$$

$$c^f(A+B+C) = c^f(C) \quad (12)$$

Applying these relations to equation (10), the next equation is obtained.

$$\begin{aligned} P &= p(c^t(A+B+C)|c^f(X)) \\ &\times p(A+B+C|c^t(A+B+C)) \\ &\times p(c^t(D)|c^f(A+B+C)) \\ &\times p(D|c^t(D)) \end{aligned} \quad (13)$$

Equation(13) means that if the frequency of the $N$ word sequence is sufficient, we can partially introduce higher order word N-grams using $N$ length word succession, thus maintaining the reliability of the estimated probability and formation of the Multi-Class 2-grams. We call Multi-Class Composite 2-grams that are created by partially introducing higher order word N-grams by word succession, Multi-Class 2-grams. In addition, equation (13) shows that number of parameters will not be increased so match when frequent word successions are added to the word entry. Only a 1-gram of word succession $A+B+C$ should be added to the conventional N-gram parameters. Multi-Class Composite 2-grams are created in the following manner.

1. Assign a Multi-Class 2-gram, for state initialization.

2. Find a word pair whose frequency is above the threshold.

3. Create a new word succession entry for the frequent word pair and add it to a lexicon. The following connectivity class of the word succession is the same as the following class of the first word in the pair, and its preceding class is the same as the preceding class of the last word in it.

4. Replace the frequent word pair in training data to word succession, and recalculate the frequency of the word or word succession pair. Therefore, the summation of probability is always kept to 1.

5. Repeat step 2 with the newly added word succession, until no more word pairs are found.

### 4.2 Extension to Multi-Class Composite 3-grams

Next, we put the word succession into the formulation of Multi-Class 3-grams. The transition probability to word sequence $(A, B, C, D, E, F)$ from word pair $(X, Y)$ is given in the next equation.

$$\begin{aligned} P &= p(c^t(A+B+C+D)|c^{f2}(X), c^{f1}(Y)) \\ &\times p(A+B+C+D|c^t(A+B+C+D)) \\ &\times p(c^t(E)|c^{f2}(Y), c^{f1}(A+B+C+D)) \\ &\times p(E|c^t(E)) \\ &\times p(c^t(F)|c^{f2}(A+B+C+D), c^{f1}(E)) \\ &\times p(F|c^t(F)) \end{aligned} \quad (14)$$

Where, the Multi-Classes for word succession $A+B+C+D$ are given by the next equations.

$$c^t(A+B+C+D) = c^t(A) \quad (15)$$

$$c^{f2}(A+B+C+D) = c^{f2}(D) \quad (16)$$

$$c^{f1}(A+B+C+D) = c^{f2}(C), c^{f1}(D) \quad (17)$$

In equation (17), please notice that the class sequence (not single class) is assigned to the preceding class of the word successions. the class sequence is the preceding class of the last word of the word succession and the pre-preceding class of the second from the last word. Applying these class assignments to equation (14) gives the next equation.

$$\begin{aligned} P &= p(c^t(A)|c^{f2}(X), c^{f1}(Y)) \\ &\times p(A+B+C+D|c^t(A)) \\ &\times p(c^t(E)|c^{f2}(C), c^{f1}(D)) \\ &\times p(E|c^t(E)) \\ &\times p(c^t(F)|c^{f2}(D), c^{f1}(E)) \\ &\times p(F|c^t(F)) \end{aligned} \quad (18)$$

In the above formation, the parameter increase from the Multi-class 3-gram is $p(A+B+C+D|c^t(A))$. After expanding this term, the next equation is given.

$$\begin{aligned} P &= p(c^t(A)|c^{f2}(X), c^{f1}(Y)) \\ &\times p(A|c^t(A)) \\ &\times p(B|A) \\ &\times p(C|A, B) \\ &\times p(D|A, B, C) \\ &\times p(c^t(E)|c^{f2}(C), c^{f1}(D)) \\ &\times p(E|c^t(E)) \\ &\times p(c^t(F)|c^{f2}(D), c^{f1}(E)) \\ &\times p(F|c^t(F)) \end{aligned} \quad (19)$$

In equation (19), the words without $B$ are estimated by the same or more accurate models than Multi-Class 3-grams (Multi-Class 3-grams for words $A$, $E$ and $F$, and word 3-gram and word 4-gram for words $C$ and $D$ ). However, for word $B$, a word 2-gram is used instead of the Multi-Class 3-grams though its accuracy is lower than the Multi-Class 3-grams. To prevent this decrease in the accuracy of estimation, the next process is introduced.

First, the 3-gram entry $p(c^t(E)|c^{f2}(Y), A+B+C+D)$ is removed. After this deletion, back-off smoothing is applied to this entry as follows.

$$\begin{aligned} & p(c^t(E)|c^{f2}(Y), c^{f1}(A+B+C+D)) \\ = \; & b(c^{f2}(Y), c^{f1}(A+B+C+D)) \\ \times \; & p(c^t(E)|c^{f1}(A+B+C+D)) \end{aligned} \quad (20)$$

Next, we assign the following value to the back-off parameter in equation (20). And this value is used to correct the decrease in the accuracy of the estimation of word $B$.

$$\begin{aligned} & b(c^{f2}(Y), c^{f1}(A+B+C+D)) \\ = \; & p(c^t(B)|c^{f2}(Y), c^{f1}(A)) \\ \times \; & p(B|c^t(B))/p(B|A) \end{aligned} \quad (21)$$

After this assignment, the probabilities of words $B$ and $E$ are locally incorrect. However, the total probability is correct, since the back-off parameter is used to correct the decrease in the accuracy of the estimation of word $B$. In fact, applying equations (20) and (21) to equation (14) according to the above definition gives the next equation. In this equation, the probability for word $B$ is changed from a word 2-gram to a class 3-gram.

$$\begin{aligned} P &= p(c^t(A)|c^{f2}(X), c^{f1}(Y)) \\ &\times p(A|c^t(A)) \\ &\times p(c^t(B)|c^{f2}(Y), c^{f1}(A)) \\ &\times p(B|c^t(B)) \\ &\times p(C|A, B) \end{aligned}$$

$$
\begin{aligned}
\times \quad & p(D|A,B,C) \\
\times \quad & p(c^t(E)|c^{f2}(C), c^{f1}(D)) \\
\times \quad & p(E|c^t(E)) \\
\times \quad & p(c^t(F)|c^{f2}(D), c^{f1}(E)) \\
\times \quad & p(F|c^t(F)) \qquad (22)
\end{aligned}
$$

In the above process, only 2 parameters are additionally used. One is word 1-grams of word successions as $p(A + B + C + D)$. And the other is word 2-grams of the first two words of the word successions. The number of combinations for the first two words of the word successions is at most the number of word successions. Therefore, the number of increased parameters in the Multi-Class Composite 3-gram is at most the number of introduced word successions times 2.

## 5 Evaluation Experiments

### 5.1 Evaluation of Multi-Class N-grams

We have evaluated Multi-Class N-grams in perplexity as the next equations.

$$Entropy = \frac{1}{N} \sum_i log_2(p(w_i)) \qquad (23)$$

$$Perplexity = 2^{Entropy} \qquad (24)$$

The Good-Turing discount is used for smoothing. The perplexity is compared with those of word 2-grams and word 3-grams. The evaluation data set is the ATR Spoken Language Database (Takezawa et al., 1998). The total number of words in the training set is 1,387,300, the vocabulary size is 16,531, and 5,880 words in 42 conversations which are not included in the training set are used for the evaluation.

Figure1 shows the perplexity of Multi-Class 2-grams for each number of classes. In the Multi-Class, the numbers of following and preceding classes are fixed to the same value just for comparison. As shown in the figure, the Multi-Class 2-gram with 1,200 classes gives the lowest perplexity of 22.70, and it is smaller than the 23.93 in the conventional word 2-gram.

Figure 2 shows the perplexity of Multi-Class 3-grams for each number of classes. The number of following and preceding classes is 1,200 (which gives the lowest perplexity in Multi-Class 2-grams). The number of pre-preceding classes is changed from 100 to 1,500. As shown in this figure, Multi-Class 3-grams result in lower perplexity than the conventional word 3-gram, indicating the reasonability of word clustering based on the distance-2 2-gram.

Table 1: Evaluation of Multi-Class Composite N-grams in Perplexity

| Kind of model | Perplexity | Number of parameters |
|---|---|---|
| Word 2-gram | 23.93 | 181,555 |
| Multi-Class 2-gram | 22.70 | 81,556 |
| Multi-Class Composite 2-gram | 19.81 | 92,761 |
| Word 3-gram | 17.88 | 713,154 |
| Multi-Class 3-gram | 17.38 | 438,130 |
| Multi-Class Composite 3-gram | 16.20 | 455,431 |
| Word 4-gram | 17.45 | 1,703,207 |

### 5.2 Evaluation of Multi-Class Composite N-grams

We have also evaluated Multi-Class Composite N-grams in perplexity under the same conditions as the Multi-Class N-grams stated in the previous section. The Multi-Class 2-gram is used for the initial condition of the Multi-Class Composite 2-gram. The threshold of frequency for introducing word successions is set to 10 based on a preliminary experiment. The same word succession set as that of the Multi-Class Composite 2-gram is used for the Multi-Class Composite 3-gram. The evaluation results are shown in Table 1. Table 1 shows that the Multi-Class Composite 3-gram results in 9.5% lower perplexity with a 40% smaller parameter size than the conventional word 3-gram, and that it is in fact a compact and high-performance model.

### 5.3 Evaluation in Continuous Speech Recognition

Though perplexity is a good measure for the performance of language models, it does not always have a direct bearing on performance in language processing. We have evaluated the proposed model in continuous speech recognition. The experimental conditions are as follows:

- Evaluation set

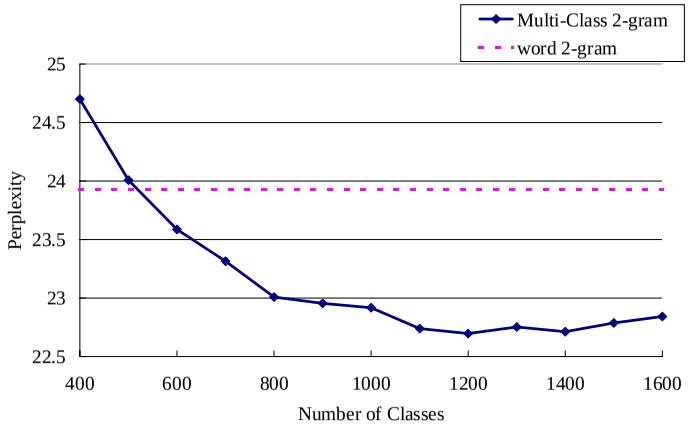


Figure 1: Perplexity of Multi-Class 2-grams

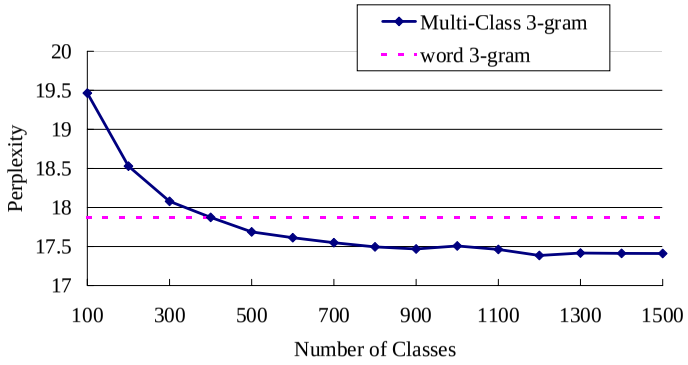


Figure 2: Perplexity of Multi-Class 3-grams

  - The same 42 conversations as used in the evaluation of perplexity

- Acoustic features
  - Sampling rate 16kHz
  - Frame shift 10msec
  - Mel-cepstrum 12 + power and their delta, total 26

- Acoustic models
  - 800-state 5-mixture HMnet model based on ML-SSS (Ostendorf and Singer, 1997)
  - Automatic selection of gender dependent models

- Decoder (Shimizu et al., 1996)
  - 1st pass: frame-synchronized viterbi search
  - 2nd pass: full search after changing the language model and LM scale

The Multi-Class Composite 2-gram and 3-gram are compared with those of the word 2-gram, Multi-Class 2-gram, word 3-gram and Multi-Class 3-gram. The number of classes is 1,200 through all class-based models. For the evaluation of each 2-gram, a 2-gram is used at both the 1st and the 2nd pass in decoder. For the 3-gram, each 2-gram is changed to the corresponding 3-gram in the 2nd pass. The evaluation measures are conventional word accuracy and %correct calculated as follows.

$$WordAccuracy = \frac{W - D - I - S}{W} \times 100$$

$$\%Correct = \frac{W - D - S}{W} \times 100$$

($W$: Number of correct words, $D$: Deletion error, $I$ : Insertion error, $S$: Substitution error)

Table 2: Evaluation of Multi-Class Composite N-grams in Continuous Speech Recognition

| Kind of Model | Word Acc. | %Correct |
|---|---|---|
| Word 2-gram | 84.15 | 88.42 |
| Multi-Class 2-gram | 85.45 | 88.80 |
| Multi-Class Composite 2-gram | 88.00 | 90.84 |
| Word 3-gram | 86.07 | 89.76 |
| Multi-Class 3-gram | 87.11 | 90.50 |
| Multi-Class Composite 3-gram | 88.30 | 91.48 |

Table 2 shows the evaluation results. As in the perplexity results, the Multi-Class Composite 3-gram shows the highest performance of all models, and its error reduction from the conventional word 3-gram is 16%.

## 6 Conclusion

This paper proposes an effective word clustering method called Multi-Class. In the Multi-Class method, multiple classes are assigned to each word by clustering the following and preceding word characteristics separately. This word clustering is performed based on the word connectivity in the corpus. Therefore, the Multi-Class N-grams based on Multi-Class can improve reliability with a compact model size without losing accuracy.

Furthermore, Multi-Class N-grams are extended to Multi-Class Composite N-grams. In the Multi-Class Composite N-grams, higher order word N-grams are introduced through the grouping of frequent word successions. Therefore, these have accuracy in higher order word N-grams added to reliability in the Multi-Class N-grams. And the number of increased parameters with the introduction of word successions is at most the number of word successions times 2. Therefore, Multi-Class Composite 3-grams can maintain a compact model size in the Multi-Class N-grams. Nevertheless, Multi-Class Composite 3-grams are represented by the usual formation of 3-grams. This formation is easily handled by a language processor, especially that requires huge calculation cost as speech recognitions.

In experiments, the Multi-Class Composite 3-gram resulted in 9.5% lower perplexity and 16% lower word error rate in continuous speech recognition with a 40% smaller model size than the conventional word 3-gram. And it is confirmed that high performance with a small model size can be created for Multi-Class Composite 3-grams.

## Acknowledgments

We would like to thank Michael Paul and Rainer Gruhn for their assistance in writing some of the explanations in this paper.

## References

Shuanghu Bai, Haizhou Li, and Baosheng Yuan. 1998. Building class-based language models with contextual statistics. In *Proc. ICASSP*, pages 173–176.

P.F. Brown, V.J.D. Pietra, P.V. de Souza, J.C. Lai, and R.L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Sabine Deligne and Frederic Bimbot. 1995. Language modeling by variable length sequences. *Proc. ICASSP*, pages 169–172.

Hirokazu Masataki, Shoichi Matsunaga, and Yosinori Sagusaka. 1996. Variable-order n-gram generation by word-class splitting and consecutive word grouping. *Proc. ICASSP*, pages 188–191.

M. Ostendorf and H. Singer. 1997. HMM topology design using maximum likelihood successive state splitting. *Computer Speech and Language*, 11(1):17–41.

Tohru Shimizu, Hirofumi Yamamoto, Hirokazu Masataki, Shoichi Matsunaga, and Yoshinori Sagusaka. 1996. Spontaneous dialogue speech recognition using cross-word context constrained word graphs. *Proc. ICASSP*, pages 145–148.

Toshiyuki Takezawa, Tsuyoshi Morimoto, and Yoshinori Sagisaka. 1998. Speech and language databases for speech translation research in ATR. In *Proc. of the 1st International Workshop on East-Asian Language Resource and Evaluation*, pages 148–155.

Hirofumi Yamamoto and Yoshinori Sagisaka. 1999. Multi-class composite n-gram based on connection direction. *Proc. ICASSP*, pages 533–536.

S. Zhang, H. Singer, D. Wu, and Y. Sagisaka. 1999. Improving n-gram modeling using distance-related unit association maximum entropy language modeling. In *Proc. EuroSpeech*, pages 1611–1614.