Multi-Class Composite N-gram Language Model Using Multiple Word Clusters and Word Successions

Shuntaro ISOGAI, Katsuhiko SHIRAI, Hirofumi YAMAMOTO, Yoshinori SAGISAKA

School of Science and Engineering, Waseda University
Okubo 3-4-1, Shinjuku-ku, Tokyo 169-8555 Japan
isogai/shirai@shirai.info.waseda.ac.jp

ATR Spoken Language Translation Research Laboratories
Hikaridai 2-2-2, Seika-cho, Soraku-gun, Kyoto 619-0288 Japan
hirofumi.yamamoto/yoshinori.sagisaka@slt.atri.co.jp

Abstract

In this paper, a new language model, the Multi-Class Composite N-gram, is proposed to avoid a data sparseness problem in small amount of 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, so-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 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

In word N-grams, the accuracy of word prediction capability will increase according to the number of the order N. However, also the number of word transition combinations will exponentially increase. Moreover, the number 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 reliable model. As a solutions for 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 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 part-of-speech (POS) word class is usually quite a bit lower than 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. In Multi-Class assignment, the word connectivity in each position of the N-grams are 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 word cluster

2.1. Class N-grams

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.

$$p(c_{u}; |c_{u-1}, \ldots, c_{u-N+1}, e_{u-1}, e_{u-2}, \ldots, e_{u-N+1})$$

Where, $c_{u}$ represents the word class to which the word $w$ belongs.

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

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 what words follow and what word precede 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. However, this word class definition cannot represent 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, different class assignment for both words will cause the information about the word connectivity will be lost. On the other hand, different class assignment for both words will cause the information about the word connectivity for what words follow and what word precede are treated as the same neighboring characteristics without distinction. Therefore, only the words that have the same 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, 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 connectivity for the words that precede and follow to be different. 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 preceding word connectivity, and to assign the different word classes to represent following word connectivity at the same time. To apply these word class definitions to formula (1), the next formula is given.

\[ p(c_v^i | y_{N+1}, \ldots, y_{N+2}, y_{N-1}, y_1) = \frac{p(w_1 | y_1) \cdots p(w_i | y_i) \cdots p(w_x | y_x)}{p(w_1 \cdots w_x | y_1 \cdots y_x)} \]  

In the above formula, \( c_v^i \) represents the word class in the target position to which the word \( w_1 \) belongs, and \( c_v^i \) represents the word class in the \( N \)-th position in a conditional word sequence.

We call this multiple word class definition, Multi-Class. Similarly, we call class N-grams based on Multi-Class, Multi-Class N-grams[6].

3. Automatic extraction of word clusters

3.1. Word clustering for Multi-Class 2-grams

For word clustering in class N-grams, sometimes POS information is used. Though POS information can be used for words that do not appear in the corpus. However, 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 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[2][1].

3.2. Word clustering for Multi-Class 3-grams

To apply the multiple clustering for 2-grams to 3-grams, the clustering of preceding word pairs are to be formulated. For 2-grams, as described in the previous section, the following words and preceding words are separately clustered. As preceding two word history for 3-grams has much wider varieties of on word history for 2-grams. Therefore, efficient word clustering is needed to keep the reliability of 3-grams after clustering and reasonable calculation cost.

Without losing the reliability caused by the date sparseness of two word history in 3-grams, and 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[7]. The vector for clustering is given in the next equation.

\[ e^{(2)}(x) = [p^{(2)}(w_1 | x), p^{(2)}(w_2 | x), \ldots, p^{(2)}(w_x | x)] \]  

Where, \( p^{(2)}(y | x) \) represents the distance-2 2-gram probability from word \( x \) to word \( y \).

4. Multi-Class Composite N-grams

4.1. Multi-Class Composite 2-gram introducing variable length word sequences

Let’s consider the condition such that only word sequence (\( A, B, C, D \)) has sufficient frequency in sequence (\( X, A, B, C, D, E \)). In this case, the value of word 2-gram \( p(B | A) \) can be used as a reliable value for the estimation of word 3, as the frequency of sequence (\( A, B, C, D \)) is sufficient. The value of word 3-gram \( p(C | A, B) \) can be used for the estimation of word 4, and word 4-gram \( p(D | A, B, C) \) can be used for word 5 for the same reason. For the estimation of words A and E, 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 (\( D, E \)) is insufficient).

Based on this idea, the transition probability of word sequence (\( A, B, C, D, E \)) from word X is given in the next equation in the Multi-Class 2-gram.

\[ P = p(c_i(A) | y_1 | A) \times p(B | A) \times p(c_i(C) | A, B) \times p(D | A, B, C) \times p(c_i(E) | D) \times p(E | c_i(E)) \]  

When word succession \( A + B + C + D \) is introduced as variable length word sequences (\( A, B, C, D \)), equation(4) can be changed exactly to the next equation[3][4].

\[ P = p(c_i(A) | y_1 | A) \times p(A + B + C + D | c_i(A)) \times p(c_i(E) | D) \times p(E | c_i(E)) \]  

Here, we find the following properties. The preceding word connectivity of word succession \( A + B + C + D \) is the same as the connectivity of word A, the first word of \( A + B + C + D \). And the following connectivity is the same as the last word D. In these assignments in equation(6) and (7), no new cluster is required. But conventional class N-grams require a new cluster for the new word succession.

\[ c_i(A + B + C + D) = c_i(A) \]  

\[ c_i(A + B + C + D) = c_i(D) \]  

Applying these relations to equation(5), the next equation is obtained.

\[ P = p(c_i(A + B + C + D) | y_1) \times p(A + B + C + D | c_i(A + B + C + D)) \times p(c_i(E) | D) \times p(E | c_i(E)) \]  

Equation(8) means that if frequency of the \( N \) word sequence is sufficient, we can partially introduce higher order word N-grams using \( N \) length words succession, thus maintaining the reliability of the estimated probability and formation of the Multi-Class 2-grams. We can Multi-Class Composite 2-grams that are partially introduced higher order word N-grams by word successions, Multi-Class 2-grams. In addition, equation(8) shows that number of parameters will not be increased so much when frequent word successions are added to the word entry. Only a 1-grams of word succession \( A + B + C + D \) should be added to the conventional N-grams parameters.
4.2. Extension to a Multi-Class Composite 3-grams

Next, we put the word succession into the formulation of Multi-Class 3-grams. Let’s consider each word transition probabilities in word sequence \((Y, X, A, B, C, D, E, F)\). In this sequence, only \((A, B, C, D)\) has sufficient frequency as the previous subsection. Before introduction of the word succession \(A + B + C + D\), the transition probabilities of the words \(A, E\) and \(F\) are given in the next formula.

\[
\begin{align*}
p(c^3(A)|r^3(Y), c^1(X)) &\cdot p(A|r^3(Y)) \\
p(c^1(E)|r^2(C), c^1(D)) &\cdot p(E|r^2(C)) \\
p(c^1(F)|r^2(D), c^1(E)) &\cdot p(F|r^2(D))
\end{align*}
\]

(9)

(10)

(11)

Next class assignments for the word succession \(A + B + C + D\) give same probabilities for the words \(A, E\) and \(F\).

\[
c^1(A + B + C + D) = c^1(A)
\]

(12)

\[
c^1(A + B + C + D) = c^2(C)c^1(D)
\]

(13)

\[
c^2(A + B + C + D) = c^2(D)
\]

(14)

After these assignments, transition probability of the word sequence \((A, B, C, D, E, F)\) from word sequence \((Y, X)\) is given next equation.

\[
P = \frac{p(c^1(A)|r^2(Y), c^1(X))p(A|r^3(Y))}{p(B|A)} \times \frac{p(c^1(E)|r^2(C), c^1(D))p(E|r^2(C))}{p(D|A, B, C)} \times \frac{p(c^1(F)|r^2(D), c^1(E))p(F|r^2(D))}{p(F|E)}
\]

(15)

In this equation, Multi-Class 3-grams are used for the words \(A, E\) and \(F\). Word 3-gram and 4-gram are used for the words \(C\) and \(D\). But, word 2-gram is used for the word \(B\). To prevent from word 2-gram to Multi-Class 3-gram, the next process is introduced.

First, the 3-gram entry \(p(c^1(E)|r^2(X), c^1(A + B + C + D))\) is removed. After this deletion, the back-off smoothing is applied for word \(E\) as the next equation.

\[
p(c^1(E)|r^2(X), c^1(A + B + C + D)) = b(c^2(X), c^1(A + B + C + D)) \times p(c^1(E)|r^1(A + B + C + D))
\]

(16)

Furthermore, back-off parameter \(b\) is defined in next equation.

\[
b(c^2(X), c^1(A + B + C + D)) = p(c^1(B)|r^2(X), c^1(A)) \times p(B|A)
\]

(17)

The equation (17) means that transition probability of word \(B\) is prevented from word 2-gram to Multi-Class 3-gram through the estimation of the word \(E\).

In the introductions of the word succeedences, only 2 parameters are additionally used. One is word 1-grams of word successions as \(p(A + B + C + D)\). And other is word 2-grams of the first 2 words of the word successions. The number of combinations for the first 2 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 in experiments

5.1. Evaluation of Multi-Class N-grams

We have evaluated Multi-Class N-grams in perplexity. The perplexity is compared with those of word 2-grams and word 3-grams. The evaluation data set is the AT&T Spoken Language Database[5]. 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.

Figure 1 shows the perplexity of Multi-Class 2-grams for each number of classes. In 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 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 (that 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 distance 2 2-gram.

5.2. Evaluation of Multi-Class Composite N-grams

We have also evaluated Multi-Class Composite N-grams in perplexity under the same condition of Multi-Class N-grams stated
Table 1: Evaluation of Multi-Class Composite N-grams in perplexity

<table>
<thead>
<tr>
<th>Kind of model</th>
<th>Perplexity</th>
<th>Number of parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word 2-gram</td>
<td>23.93</td>
<td>181,555</td>
</tr>
<tr>
<td>Multi-Class 2-gram</td>
<td>22.70</td>
<td>81,356</td>
</tr>
<tr>
<td>Composite 2-gram</td>
<td>19.81</td>
<td>92,761</td>
</tr>
<tr>
<td>Word 3-gram</td>
<td>17.88</td>
<td>713,154</td>
</tr>
<tr>
<td>Multi-Class 3-gram</td>
<td>17.38</td>
<td>438,130</td>
</tr>
<tr>
<td>Composite 3-gram</td>
<td>16.20</td>
<td>455,431</td>
</tr>
</tbody>
</table>

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 speech recognition. We have evaluated the proposed model in continuous speech recognition under the same conditions as the perplexity.

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

\[
\text{Word Accuracy} = \frac{W - D - I + S}{W} \times 100
\]

\[
\% \text{Correct} = \frac{W - D - I}{W} \times 100
\]

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

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, multiple classes are assigned to each word by clustering following and preceding word characteristics separately. These word clustering is performed based on the word connectivity in the corpus. The Multi-Class N-grams based on Multi-Class can improve reliability with a compact model size without losing their accuracy.

Next, 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.

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. It is confirmed that high performance with a small model size can be created for Multi-Class Composite 3-grams.

7. References


