AN EVALUATION OF STATISTICAL LANGUAGE MODELING FOR SPEECH RECOGNITION USING A MIXED CATEGORY OF BOTH WORDS AND PARTS-OF-SPEECH

Yumi Wakita, Jun Kawai, Hitoshi Iida

ATR Interpreting Telecommunications Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

ABSTRACT

In our previous paper, we proposed a mixed category of words and parts-of-speech names the MWP category based on class N-gram modeling [1]. However, we had not confirmed the efficiency of MWP category. In this paper, we evaluate the proposed MWP category. At first we use "coverage of words and category sequences to open data" and "perplexity to training data" for the evaluation and we confirmed the characteristics of parts-of-speech are useful to for generating a suitable class N-gram modeling. As a result of the speech recognition experimentation, we also confirmed that the class N-gram modeling using MWP category is effective in improving the recognition rate for open data that shows a low coverage of words and category sequences, without decreasing the recognition rate much for closed data.

1. INTRODUCTION

A statistical language modeling called N-grams has been widely used for continuous speech recognition. The effectiveness of N-grams depends on the relation between the training data and test data. When the test data includes words or the word sequences unknown in training data, the effectiveness of N-grams decreases much.

Smoothing techniques [2][3] and class-based modeling [4] have been proposed to efficiently describe a corpus with a limited number of training corpora. However, smoothing techniques are unable to assign reliable probabilities to low appearance frequency words and word sequences, and class-based modeling is unable to predict the exact connection of words.

We proposed a mixed category of words and parts-of-speech named the MWP category based on class n-gram modeling for giving a reliable probability to both high and low appearance frequency words and category sequences [1]. We guess that the MWP category can solve conventional n-gram problems, because of "parts-of-speech categories" can provide reliable probabilities to low appearance frequency words and the "word categories" can predict the exact connection of words.

Another N-gram modeling using mixed categories of parts-of-speech and words are proposed [5]. In this method, the parts-of-speech categories are split into word categories by minimizing the total entropy of the training data. The method can reduce perplexity most suitably for the training data , but the effect depends on the training corpus.

In the proposed MWP category a category is defined using the characteristics of parts-of-speech. Categories are defined by each parts-of-speech unit. Any two words that belong to the same parts-of-speech are always given the same category tag.

In this paper we evaluate the proposed MWP category using the "coverage to open data" and the "perplexity of training corpus". As a result of evaluation, it is found that the characteristics of parts-of-speech is useful in obtaining a suitable MWP category. As a result of recognition experimentation, the proposed MWP category is effective in improving the recognition rate for open data, without decreasing the recognition rate much for closed data.

2. MIXED CATEGORY OF BOTH WORDS AND PARTS-OF-SPEECH

The probability $Pr(W,T)$ where W is the word sequence occurrence and boldmath T is the category sequence occurrence is represented by probability theory as:

$$Pr(W,T) = Pr(T)Pr(W|T) \tag{1}$$

For an n-gram class model we have:

$$Pr(W,T) \approx \prod_{i=1}^{N} Pr(t_i|t_{i-n+1}^{i-1})Pr(w_i|t_i) \tag{2}$$

where $Pr(t_i|t_{i-n+1}^{i-1})$ is the probability that category t_i will occur given that category sequence $\{t_{i-n+1}, t_{i-n+2}, \cdots, t_{i-1}\}$ has occurred previously, and $Pr(w_i|t_i)$ is the probability that word w_i occurs in t_i.

An MWP category tag, $t_m(w)$, which a word w belongs to, is defined as follows:

$$t_m(w) = \begin{cases} POS(w), & \text{if } w \in W_{indep} \\ w, & \text{if } w \in W_{dep} \end{cases} \tag{3}$$

Table 1: Experimental conditions 1

n-grams	bi-gram of MWP category
task	speech conversation by 'clerk' and 'customer' on travel arrangement
bi-gram training	customer's 3363 sentences included 37159 words
number of POS	27
test data	(closed) 145 sentences in training data, 10 customer (open1) 107 sentences not in training data, 10 customer (open2) 166 sentences not in training data, 10 clerk

where $POS(w)$ is the parts-of-speech which the word w belongs to, and W_{indep} is a category including word w that satisfies the following conditions

cond1 word w belongs to the parts-of-speech that include a lot of high frequency words independent of the corpus.

cond2 word w belongs to the parts-of-speech including a lot of words that have no singularity in word-to-word connections.

W_{dep} is a category including the word w which does not satisfy the above conditions.

In our previous paper [1], we stated that function words belong to word categories, and content words belong to parts-of-speech categories. However we didn't confirm the efficiency of these categories.

In the next section, we describe an evaluation of MWP categories using coverage to open data and perplexity to training data.

3. EVALUATION OF MWP CATEGORY USING COVERAGE AND PERPLEXITY

To define the category so that the parts-of-speech category satisfies the two conditions shown in section 2, we aim at N-grams using MWP category that shows high coverage to open data and low perplexity to closed data. The reasons for using 'coverage' and 'perplexity' are as follows:

- Perplexity shows the average number of predicted words. In particular, the prediction ability of connection of words can be estimated by using perplexity to closed data. Coverage to open data is useful in estimating bad effects due to unknown words and category sequences. These factors are useful for estimating the effective MWP category directly.
- Coverage and perplexity affect the recognition rate with statistical language modeling. These factors are useful in predicting the recognition performance exactly.

To confirm that the parts-of-speech unit is suitable for the MWP category, we counted the number of occurrence of unknown words in open data and sorted them according to each parts-of-speech, and we calculated the coverage and the perplexity of the MWP category. The experimental conditions are shown in Table1 The bi-gram probability was calculated using 3369 sentences (37159 words) spoken by the "client" : the group names of the parts-of-speech are as follows; a number in brackets shows how many different kind of parts-of-speech there are in a group, noun (7), pronoun (2), verb (3), auxiliary verb (3), adjective (2), adverb (2), conjunction (1) and post-positional particle (7).

3.1. Coverage for open data by each part-of-speech

We counted up the number of known words and the category sequences to the language model for the open data. We used two sets of open data, one spoken by client under the same condition as for the training data, and the other spoken by the clerk a different condition from the training data. The coverage of the language model for the test set was calculated using the following formula:

$$Coverage = \frac{\sum_{i}^{N} f_{known}(w_{i-n+1}^{i})}{N} \quad (4)$$

where N is the number of all word sequences, w_{i-n+1}^{i}, appearing in the test set. f_{known} is a binary function and defined as follows:

$$f_{known}(w_{i-n+1}^{i}) = \begin{cases} 1, & \text{if both } t_{i-n+1}^{i-1} \text{ and } w_i \\ & \text{is known to} \\ & \text{the language model.} \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

where t_{i-n+1}^{i-1} is the category sequence for the word sequence w_{i-n+1}^{i-1} and w_i is the last word of the word sequence w_{i-n+1}^{i}.

Table2 is an example showing parts-of-speech which include a lot of unknown words. We confirmed the following:

- Over 80% of unknown words are included in common-noun or verb or nouns followed by "suru". Almost unknown words appear in limited parts-of-speech.
- The tendency of the appearance of unknown words for the client's data is the same as clerk's data.

These result shows that the parts-of-speech unit is useful for dividing words into low appearance frequency words and high appearance frequency words independently of corpus.

Table 2: Example of parts-of-speech including many unknown words (percentage of unknown words in each part-of-speech to all unknown words)

customer	clerk
common noun (45.0%)	common noun (53.1%)
verb (20.0%)	verb (22.4%)
nouns followed by "suru"(15.0%)	nouns followed by "suru"(5.1%)
adjective noun (10.0%)	auxiliary verb (3.1%)

Table 3: Data conditions for recognition experiment

acoustic model	speaker independent HMnet; 401 states, 10 mixtures per state
language model	bi-gram of MWP category
recognition	one-pass DP, N-best search
task	speech conversation between a 'clerk' and 'client' on travel arrangements
training corpus for bi-gram	client's 3363 sentences included 222954 words
number of POS	27
test data	(closed) 145 sentences in the training corpus, 3 clients (open1) 107 sentences not in the training data, 3 clients (open2) 166 sentences not in the training data, 2 clerks

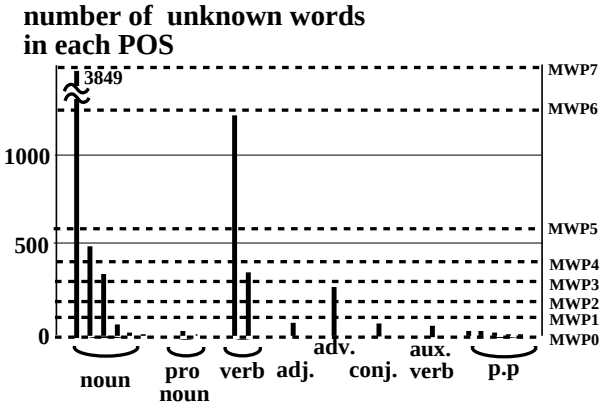

Figure 1: Boundary between POS and word category of each n-gram using MWP category

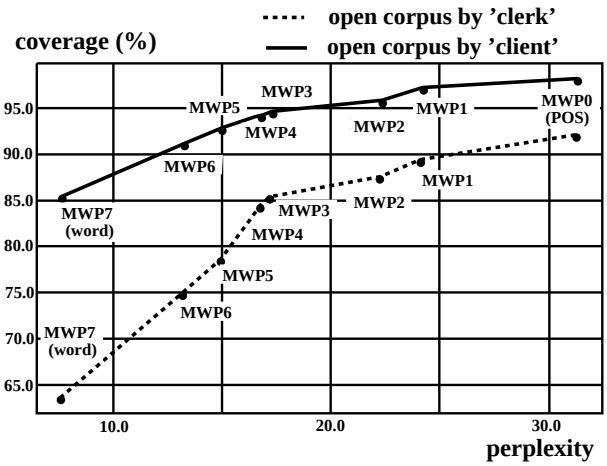

Figure 2: Relation between coverage and perplexity

3.2. Relation between coverage and perplexity

The result of the previous experimentation show that it is suitable to use the parts-of-speech unit to design an MWP category that shows a high coverage. Next we examine the possibility of designing an MWP category shows a low perplexity.

To examine this possibility, we generated several MWP categories in the following way and examined the perplexities of these categories. At first all categories were set to parts-of-speech as initial categories. Various levels of the MWP categories were generated by changing categories from parts-of-speech to words one by one, and the parts-of-speech showing a higher coverage were changed earlier.

We selected seven MWP categories from among various MWP categories for experiments. Figure1 shows the relation between the number of unknown words in each parts-of-speech and seven MWP categories. The number of unknown words are counted using the open data which included 5280 sentences spoken by the "clerk" The vertical unbroken lines mean the number of unknown words sorted to each parts-of-speech. The horizontal dotted lines are boundaries between the parts-of-speech and the words for each categories (MWP0-MWP7). For example if a parts-of-speech include more unknown words than the boundary of "MWP3", the category of the words in this parts-of-speech is "parts-of-speech" for the MWP3 category, and if a parts-of-speech include less unknown words than the boundary of "MWP3", the category of the words in this parts-of-speech is "word" category. The words in pronoun or adjective or adverb group are included "word" category and the word in some noun and verb group are included "parts-of-speech" category. MWP0 means all categories are "parts-of-speech" and MWP7 means all categories are "word".

Figure2 shows the relation between the coverage for open

data and the perplexity for closed data.

We confirmed the following things:

As it is natural that a lower coverage shows a lower perplexity. However, the rate of decrease is different for each part-of-speech. For example a "Post-positional particle" or "auxiliary verb" include few unknown words and shows a large decreasing in of perplexity(MWP0 - MWP3), but some parts-of-speech in "noun" or "verb" group show a large decrease in coverage (MWP4 - MWP7) .

When the category change from "MWP3" to "MWP4", the slopes change sharply. this tendency is shown by both "clerk" and "client".

These results show that it is possible to design an MWP category that gives a high perplexity and a low coverage. When the category is defined to "MWP3", the balance between coverage and perplexity is the best for both of two corpus.

4. RECOGNITION EXPERIMENT

We confirmed the efficiency of the proposed MWP category by speech recognition experiments. We compared the recognition rate using several levels n-grams of MWP categories; those are used in section 3. The data conditions are shown in Table 3. We don't use usual smoothing techniques to evaluate the effective of the MWP category purely. The average recognition rate per speaker in categories "POS" and "MWP3" and "word" are shown in Figure 3. The horizontal axis means perplexity value and the value in brackets are the perplexity of each MWP category. The values in brackets shows the perplexity,

As it is easy to predict, in the case of the closed test set, "word" shows best score (88.0%) while the score by "MWP3" decreases about 10%. But the difference is smaller than the difference of between "POS" and "MWP3". In the case of test set "open1" spoken by client (same as training data), "word" shows the best score, but the difference of "word" and "MWP3"(70.9% - 68.7%) is a very small. In the case of test set "open2" spoken by clerk (different from training data), "MWP2" shows the best score (53.8%).

The proposed MWP category is more effective for the test set that coverage is lower. The recognition rate for MWP category is almost the same rate for word category by using the case of "open1" set ,that coverage of words is 85.4%. In this experimentation, the proposed MWP category is effective for test data which coverage of words is under 85%.

These results show that MWP categories can improve the recognition rate for open data that show low word coverage, without decreasing the recognition rate much for closed data.

5. CONCLUSION

We evaluated the mixed category of both words and parts-of-speech generated by using the characteristics of parts-of-speech. By using 'coverage' and 'perplexity', we confirmed that the characteristics of parts-of-speech are useful for generating a suitable class n-gram modeling. we also confirmed by speech recognition experimentation that the proposed MWP categories are effective in improving the recognition rate for open data that shows the low coverage of words and category sequences, without decreasing the recognition rate much for closed data. From these evaluation, We expect the factors of coverage and perplexity are useful to generate the optimum MWP category for both of closed and open data. In future work, we'll study the method to generate the optimum MWP category.

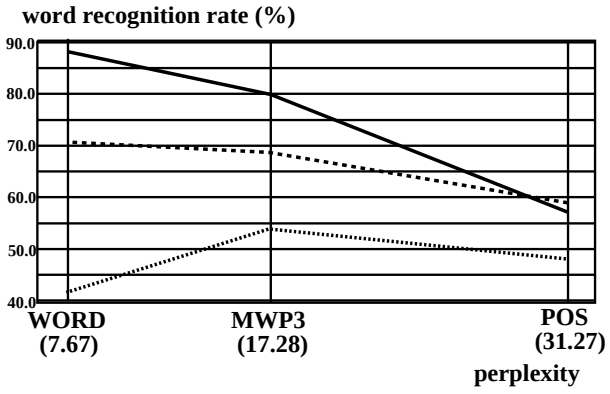

Figure 3: Comparison of various N-gram modeling

ACKNOWLEDGMENTS

The authors wish to acknowledge the help received from Dr. Yamazaki, resident of ATR Interpreting Telecommunications Research Laboratories.

REFERENCES

[1] Kawai, J., Wakita, Y. and Iida, H.: Stocastic language model using semantic category and mixed category of words and parts-of-speech for speech understanding, *Proc. NLPRS95*, Vol.1, pp.107-111, 1995.

[2] Jelinek, F. and Mercer, R.L.: Interporated estimation of Markov source parameters from sparse data, *Proc. of the Workshop on Pattern Recognition in Practice*, Amsterdam, The Netherlands, North-Holland, 1980

[3] Katz, S. K.: Estimation from sparce data for the language model for a speech recognition, *IEEE Trans. ASSP-35*, 3, pp.400-401, 1987

[4] Brown, P. F., Della Pietra, V. J., deSouza P. V., Lai, J. C., and Mercer, R. L.: Class-Based *n*-gram Models of Natural Language, *Computational Linguistics*, Vol.18, No.4, pp.467-479, 1992.

[5] Masataki, H. and Sagisaka Y.: Valiable-Order N-gram Generation by Word-Class Splitting and Consecutive Word Grouping,*Proc. ICCASP 96*