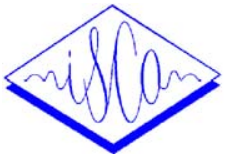

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

PART-OF-SPEECH N-GRAM AND WORD N-GRAM FUSED LANGUAGE MODEL

Hirofumi Yamamoto and Yoshinori Sagisaka
ATR Interpreting Telecommunications Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
yama@itl.atr.co.jp

ABSTRACT

In this paper, an accurate and compact language model is proposed to cope robustly with data sparseness and task dependencies. This language model adopts new categories which are generated by continuously interpolating POS word-class categories and word categories using MAP estimation. The new categories can reflect word statistics efficiently without loosing accuracy and task-independent general word-characteristics (i.e. grammatical constraints captured by POS statistics) are embedded to prevent task-overtuning. This modeling reduces the model size to 50% of the conventional models. The bi-directional word-cluster N-grams generated by this modeling have 3% lower perplexity measured on a matched domain and 15% lower on a mismatched domain compared to a conventional word 2-gram.

Keywords: Word clustering, Class N-gram, MAP Estimation

1. INTRODUCTION

N-gram language models are widely used in LVCSR. They are expected to have the following four properties:

1. accurate word prediction capability,
2. reliability for sparse data,
3. compact model size,
4. robust for different task domains.

Word N-grams are quite accurate for the prediction of the next word, but not reliable for sparse data and not robust for different task domains. Class N-grams based on part-of-speech (hereafter referred to as POS class N-gram) are better than word N-grams for the above (2) ~ (4) properties, but significantly inferior to the word N-gram in terms of word prediction accuracy. To cope with this prediction inferiority, automatic classification algorithms were proposed[3][4]. The automatically genereted class 2-gram models show good performance on sparse data and lead to compact models. However, they are heavily the task dependent.

In this paper, we propose a new N-gram model which can provide characteristics better than the conventional ones in above four properties. The new model is robust to task mismatch and reliability for sparse data. The new N-grams are calculated by estimating the maximum a-posteriori probability of the word 2-gram using the POS class 2-gram as a-priori knowledge. Word grouping and clustering are carried out based on these new N-gram statistics to attain a compact and robust language model without losing accuracy and reliability.

2. WORD FREQUENCY DEPENDENT INTERPOLATION OF N-GRAM STATISTICS BETWEEN WORD AND POS

2.1 Interpolation by MAP Estimation

The reliability of word pair statistics of the word 2-gram depends on the frequency of occurence of each word pair. If we can use the information from the POS class 2-gram for rarely occurring word pairs, we expect to increase the reliability without decreasing the accuracy while maintaining the robustness of the class 2-gram against mismatch of domain. This idea can be realized in the following ways according to the frequency-of-occurence: (a) switch from word 2-grams to POS class 2-grams[4], (b) linearly combine them [1], or (c) use backoff smoothing[5]. However, statistical reliability is uniformly assumed in these method. As word frequency differs greatly from word to word, it is expected to be much more accurate and reliable to interpolate word and POS statistics in proportion to word frequencies. We use MAP estimation for this interpolation. POS class 2-grams are used as a-priori probabilities and word 2-grams as a-posteriori probabilities.

2.2 Estimation of A-priori Distribution

The word transition probability for each word pair (w_i, w_j) is represented by the transition probability of the POS 2-gram

$$p_{class}(w_j \mid w_i) \cong p(c_j \mid c_i)\,p(w_j \mid c_j) \qquad (1)$$

where (c_i, c_j) is the class pair that the word pair (w_i, w_j) belongs to. The a-posteriori probability is high that the value of the word 2-gram is close to the value of th POS class 2-gram. And the probability is low that their values are far from each other. We assume that the distribution representing this probability, i.e. a-priori distribution is a Beta-distribution.

$$\frac{1}{B(\alpha,\beta)} p^{(\alpha-1)}(1-p)^{(\beta-1)} \quad (2)$$

For POS to word 2-gram, we make the same assumption. Let $C(w_i, w_j)$ be the measured frequency of word pairs (w_i, w_j), and $C(w_i)$ be an occurrence of word w_i. The transition probability from word w_i to word w_j can be expressed by the following equation through MAP estimation.

$$p_{MAP}(w_j \mid w_i) = \frac{C(w_i, w_j) + \alpha - 1}{C(w_i) + \alpha + \beta - 2} \quad (3)$$

From the assumption of Beta-distribution, its mean and variance are given as follows.

$$\mu = \frac{\alpha}{\alpha + \beta} \quad (4)$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \quad (5)$$

However it is difficult that we obtain α and β in maximum likelihood estimation. When $C(w_i)$ and $C(w_i, w_j)$ is 0, (i.e. the a-posteriori knowledge can not be obtained,) it is reasonable to use the a-priori knowledge. We get the next equation by applying $C(w_i) = C(w_i, w_j) = 0$ in Eq. (3).

$$p_{MAP}(w_j \mid w_i) = \frac{\alpha - 1}{\alpha + \beta - 2} \quad (6)$$

When the a-posteriori knowledge can not be obtained, $p_{MAP}(w_j \mid w_i)$ is equal to the a-prior value.

$$p_{MAP}(w_j \mid w_i) = p_{class}(w_j \mid w_i) = \hat{\mu} \quad (7)$$

Here, $p_{MAP}(w_j \mid w_i)$ is $\hat{\mu}$, a uni-gram weighted mean of observation values at the same time. The relation of μ in Eq. (4) and $\hat{\mu}$, α is substituted by α -1, and $\alpha+\beta$ is substituted by $\alpha+\beta-2$, whitch is easily solved for α and β. We get next equation by applying this replacement to Eq. (5)

$$\hat{\sigma}^2 = \frac{(\alpha-1)(\beta-1)}{(\alpha+\beta-2)^2(\alpha+\beta-1)} \quad (8)$$

Here, $\hat{\sigma}^2$ is the uni-gram weighted variance of observation values. α and β are obtained as follows from Eq. (6) and (8).

$$\alpha - 1 = \frac{\hat{\mu}^2}{1-\hat{\mu}}\hat{\sigma}^2 - \hat{\mu} \quad (9)$$

$$\alpha + \beta - 2 = \frac{\hat{\mu}(1-\hat{\mu})}{\hat{\sigma}^2} - 1 \quad (10)$$

MAP estimation is performed for each word pair by applying Eq. (3) using these α and β.

2.3 Normalization and Smoothing

Since $p_{MAP}(w_j \mid w_i)$ are calculated independently for each word pair, the $p_{MAP}(w_j \mid w_i)$ for each word may require normalization to make their sum become 1. After MAP estimation, we can not assign probability value to word pairs that are unseen in a-priori knowledge. To these word pairs, we assigned probability values by using back-off smoothing. After normalizing and smoothing, $p_{MAP}(w_j \mid w_i)$ values are obtained by the following algorithm:

If $C(w_i, w_j) \neq 0$

$$p(w_j \mid w_i) = d(w_i)\frac{C(w_i, w_j) + \alpha - 1}{C(w_i) + \alpha + \beta - 2} \quad (11)$$

If $C(c_i, c_j) \neq 0$

$$p(w_j \mid w_i) = d(w_i)\frac{\alpha - 1}{C(w_i) + \alpha + \beta - 2} \quad (12)$$

If $C(c_i, c_j) = 0$

$$p(w_j \mid w_i) = b(w_i)\,p(w_j) \quad (13)$$

where $d(w_i)$ is the discount coefficient, and $b(w_j)$ is the back-off coefficient.

As seen in Eq. (3), the proposed 2-gram values are close to word 2-gram values for frequent word pairs and close to POS class 2-gram for infrequent word pairs.

3. EVALUATION OF POS-INTERPOLATED 2-GRAM

3.1 Condition of Experiment

The proposed POS-interpolated word 2-grams are evaluated in comparison with conventional 2-gram such as word 2-grams, POS class 2-grams, and linear combinations of them. The training set consists of 260,000 words with 4,000 unique words; the a-posteriori knowledge for the MAP estimation is a word 2-gram, and the a-priori knowledge is a multi-class 2-gram [4] based on POS information from 75 To-classes and 88 From-classes.

In the multi-class 2-gram, transition probabilities from word w_i to word w_j is given as:

$$p(w_j \mid w_i) \cong p(ct_j \mid cf_i)\,p(w_j \mid ct_j) \quad (14)$$

Where, cf_i is the From-class that represents following word connectivity characteristics of word w_i, and ct_j is the To-class that represents preceding word connectivity characteristics of word w_j. From-class and To-class can be defined independently.

Test set A has 16 conversations with 2,178 words from the same domain as the training data. Test set B consists of 24 conversations with 3,655 words from a domain different from the training data.

3.2 Results

The perplexities for each model for test set A and B are shown in Table 1. It has turned out that the proposed MAP estimated 2-gram with POS classes (f), (g) show the lowest perplexity for each testset, especially for different tasks. The perplexity of the model using the multi-class 2-gram as a-priori knowledge (f) is a little higher than that of the model using the class word 2-gram as a-priori knowledge (g). This difference seems to result from the following fact. POS class is a rough classification to express word statistics. In the linear combination model, the perplexity of the model with the multi-class 2-gram (d) is higher than with POS to word 2-gram (e).

Table 1. Comparison of perplexity by each model

	Test Set A (Same Task)	Test Set B (Differnt Task)
(a) Word 2-gram	14.39	53.06
(b) Part-of-Speech Class 2-gram	30.97	74.70
(c) Class-word 2-gram	22.00	58.74
(d) Liner 2-gram (with POS Class)	14.21	46.61
(e) Liner 2-gram (with POS-Word)	14.09	45.69
(f) MAPed 2-gram (with POS Class)	13.75	47.04
(g) MAPed 2-gram (with POS-Word)	13.33	43.76

4. CLUSTERING ACCORDING TO POS-INTERPOLATED 2-GRAM STATISTICS

4.1 Feature values in Automatic Clustering

When we use the word 2-gram values as word features for clustering, the obtained classes will be task dependent, since the feature values depend on the task. The proposed MAP estimated 2-gram with POS classes are robust to task mismatch. So we can expect that clustering with MAP estimated 2-gram values as word features give us classes witch are robust against task mismatches. When we use the multi-class model as in Eq. (14), we can classify the From-class and the To-class independently. From- class presents following word connectivity characteristics,. We can use forward MAP estimated 2-grams with POS to class 2-grams, because forward connectivity characteristices are interpolated by POS statistics. In the same way we can use backward MAP estimated 2-grams for the To-class. Here, after model using this clustering, is refered to as fused 2-gram with POS class 2-gram and word 2-gram.

4.2 Procedure of Clustering

The automatic classification is performed as follows:

1. Make one class per word
2. For each word X, assign these values as features:

$$V(X) = V_t(X) \quad (15)$$

$$V(X) = V_f(X) \quad (16)$$

where $V_t(X)$ is the 2-gram after MAP estimation of the vector from X backward, and $V_f(X)$ is the same for the vector from X forward.

3. Choose the class pair which minimizes the cost function Unew - Uold, and merge these into one class. U_{new} and U_{old} are given by the next equation.

$$U_{new} = \sum_w p(w) D(V(C_{new}(w)), V(w)) \quad (17)$$

$$U_{old} = \sum_w p(w) D(V(C_{old}(w)), V(w)) \quad (18)$$

C_{new} is the class after merging, C_{old} is the class before merging, and D(vc,vw) represents the square of the Euclidean distance between vector vc and vw.

4. Repeat step 2 and 3 to reduce the number of classes as needed.

5. EVALUATION OF FUSED N-GRAM

5.1 Experimental Conditions and Results

The effect of word clustering according to POS-interpolated 2-gram statistics is evaluated and compared to using word 2-grams and automatic class 2-grams. The conditions of training and test set are the same as in section 3. Figure 1 shows the effect of word clustering among test sets A, B. Perplextity decreases as the number of classes decreases, especially in set B (in different task). with the conventional model. The fused 2-gram with 500 classes shows lower perplexity than any of the conventional model. Especially in test set B, that is different task of training set, for the word 2-gram, the perplexity decreases by 15%. The logical parameter size of the fused 2-grams is only 2% of the conventional word 2-gram and the number of entries is half. These results confirm that the fused 2-gram gives us compact models that are robust against task mismatches. Table 2 shows the comparison of fused 2-grams

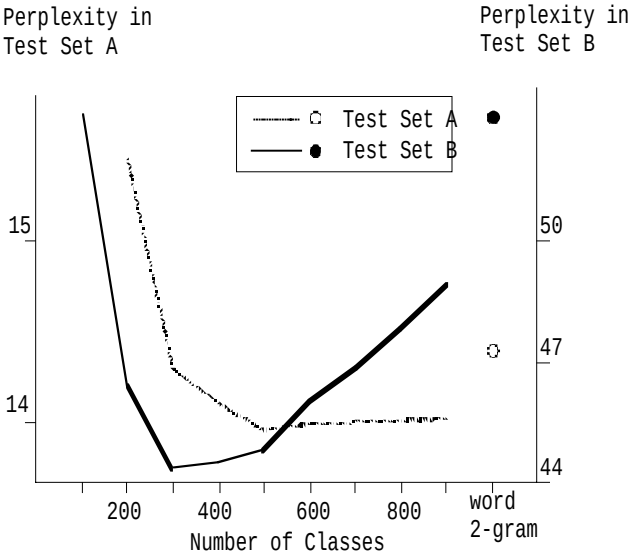

Figure 1. Effect of decreasing of perplexity by word clustering

Table 2. Performance of Fused 2-gram in Perplexity

	Number of Classes	Test Set A (Same Task)	Test Set B (Different Task)
Word 2-gram		14.39	53.06
Automatic Class 2-gram	500	15.12	59.10
	200	16.62	58.01
Fused 2-gram	500	13.95	45.32
	200	15.44	46.51

6. CONCLUSION

A POS-interpolated N-gram using MAP estimation is proposed. Futhermore, a fused N-gram is proposed by clustering words according to POS-interpolated N-gram statistics. While maintaining the prediction accuracy of the word 2-gram, this model has 2% of the logical parameters, and 50% of the entries of word 2-gram. The proposed models are reliable with sparse data and are robust against task mismatches, with a perplexity 3% lower on the same task, and 15% lower on different tasks.

7. ACKNOWLEDGMENTS

We'd like to thank Ben Reaves for assistance in writing some of the explanations in this paper.

8. REFERENCES

[1] F. Jelinek and R. L. Mercer: "Interpolated Estimation of Markov Source Parameters from Sparse Data," Pattern Recognition in Practice, pp. 381-397 North Holland, 1980

[2] H. Masataki, Y. Sagisaka, K. Hisaki and T. Kawahara: "Task Adaptation Using MAP Estimation in N-gram Language Modeling," Proc ICASSP, vol 1, pp. 783-786, 1997

[3] Shuanghu Bai, Haizhou Li, Zhiwei Lin and Baosheng Yuan: "Building Class-based Language Models with Contextual Statistics," Proc ICASSP, vol. 1, pp. 173-176, 1998

[4] H. Yamamoto and Y. Sagisaka: "Multi-Class Composite N-gram Language Model Based on Connection Direction," Proc ICASSP, vol. 1, pp. 533-536, 1999

[5] C. Samuelsson and W. Reichl: "A Class-Based Language Model for Large-Vocabulary Speech Recognition Extracted from Part-of-Speech Statistics," Proc ICASSP, vol 1, pp. 537-540, 1999