BALANCING ACOUSTIC AND LINGUISTIC PROBABILITIES

Atsunori OGAWA, Kazuya TAKEDA and Fumitada ITAKURA

Dept. of Information Electronics
Nagoya University
NAGOYA, 465 JAPAN

ABSTRACT

The length of the word sequence is not taken into account under language modeling of n-gram local probability modeling. Due to this property, the optimal values of the language weight and word insertion penalty for balancing acoustic and linguistic probabilities is affected by the length of word sequence. To deal with this problem, a new language model is developed based on Bernoulli trial model taking the length of the word sequence into account. Not only better recognition accuracy but also robust balancing with acoustic probability compared with the normal n-gram model of the proposed method is confirmed through recognition experiments.

1. INTRODUCTION

A merit of stochastic approaches for continuous speech recognition is not only its self-organizing property, but also the ease of combining two or more different knowledge sources [**?**]. The basic formula of speech recognition, for example, has a simple form of multiplying two probabilities given by acoustic knowledge $P(A|W)$ and linguistic knowlege $P(W)$, i.e.

$$\log P(A,W) = \log P(A|W) + \log P(W),$$

for combining two different knowledge sources. However, in practice, the simple multiplication needs to be modified for balancing the absolute values of two probabilities [**?**],[**?**]. This is because the two values are not true probabilities but approximations.

The most common modification for balancing two probabilities is to use a language weight α and insertion penalty Q, i.e.

$$\log \hat{P}(A,W) = \log P(A|W) + \alpha\{\log P(W) - nQ\},$$

where n is the word length of the sequence W. (An alternative formulataion $\log P(A|W) + \alpha \log P(W) - nQ$ is adopted in some systems such as HTK[**?**].) Since the probabilistic meaning of these two parameter is not clear, both of them are determined experimentally.

The purpose of this paper is to discuss and solve the three problems of this heuristic balancing; 1) the balancing parameters are critical to recognition accuracy; 2) the optimal values of two parameters are related with each other; and 3) the optimal values are also governed by the length of word sequence. In the rest of this paper, we first present the above mentioned problems from experimental results. Then, we propose Ngram-Bernoulli language modeling, which is a simple extension of Ngram model but takes the length of the word sequence into account, for removing the sequence length dependency from the combining process. Finally, we will show the effectiveness of the proposed modeling by experimental results.

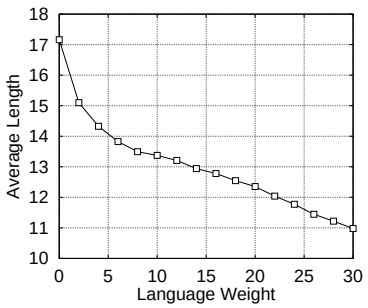
2. PROBLEMS IN COMBINING PROBABILITIES

In n-gram language modeling, the probability of an L-word-long partial word sequence is given as the L-time product of the local probabilities. The simple consequence of this modeling is that a lower probability is assigned to a longer word sequence, i.e.

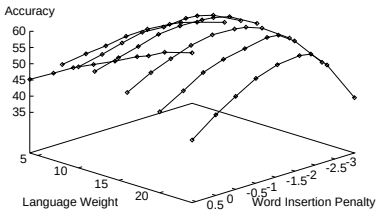
$$P(\mathbf{w}_1^n) \geq P(\mathbf{w}_1^{n+1}).$$

For combining with acoustic probability, a problem arises due to this simple decreasing property of n-gram. When the language probability is weighted too heavily, due to preference of shorter word sequences in n-gram modeling, the output of the recognizer is likely to be shorter than the correct answer. Thus the optimal value of α depends on the word-length of the sentence to be recognized. Experimental evidence of this shortening is shown in Figure 1 (a) as the averaged length of recognition results versus α.

Word Insertion Penalty (WIP) is used as another heuristic in order to compensate for the word length dependency of ngram modeling, as follows. When the language probability is not weighted enough, or α is optimized for shorter sentences, giving a penalty propotional to the word length helps to inhibit word insertions. Instead, when language probability is weighted too heavily, negative WIP works to inhibit word deletion errors. Thus we have to optimize α for



(a) Monotonic decreasing of the averaged word length of recognition results.



(b) Dependency between α and Q in potimizing.

Figure 1: Problems of heuristic parameters in combing acoustic and linguistic probabilities. Recognition experiment results using bigram LM.

the difference in scale of acoustic and linguistic probabilities and then optimize Q for the particular α. This is also confirmed experimentally as shown in Figure 1 (b).

3. LANGUAGE MODELING TAKING SENTENCE LENGTH INTO ACCOUT

3.1. Unigram-Bernoulli modeling

In order to incorporate the length of the word sequence in modeling, we start with generalized Bernoulli trials using unigram probability as the probability of event occurence,

$$P_{\text{UGBER}}(\mathbf{W}) = \frac{n!}{\prod_{j=1}^{L} k_j(\mathbf{W})!} \prod_{j=1}^{L} \{P(w_j)\}^{k_j(\mathbf{W})}, \qquad (1)$$

where $k_j(\mathbf{W})$ is the frequency of word w_j in the word sequence $\mathbf{W}$, n is the word length of the sequence, L is vocaburary size and $P(w_j)$ is the unigram probability. From now on we call the model Unigram-Bernoulli language model.

In the model first term $n!/\prod_{j=1}^{L} k_j!$ accounts for the number of combinations of word sequences constrained by the frequency of each word. Note that if each word appears only once, the term becomes $n!$, which represents the total number of possible orders of word sequences. The second term $\prod_{j=1}^{r} \{P(w_j)\}^{k_j(\mathbf{W})!}$ calculates the probability of each ordered word sequence.

3.2. Ngram-Bernoulli modeling

Unigram-Bernoulli model takes sentence length into account, however, unigram does not provide any constraint concerning word order. Thus, we try to replace unigram with n-gram local probabilities. In n-gram cases, the event space becomes a Cartesian product of word occurences. In the bigram case, for example, each trial is associated with word occurrence conditioned by the previous word. This can be realized by assigning $P(w_i|w_{i-1})$ for event probability, word length n for number of trials and occurrence of each word pair in the sequence as k_j, i.e.

$$P_{\text{BGBER}}(\mathbf{W}) = \frac{n!}{\prod_{l=1}^{L}\prod_{m=1}^{L} k_{i|i-1}(l|m)!} \prod_{i=1}^{n} P(w_i|w_{i-1}). \qquad (2)$$

From the above mathematical formula, it is clear that the Ngram-Bernoulli model is a sequence length dependently weighted version of the conventional ngram language models. Thus, although the formulation started with the assumption of independent trials, ngram-Bernoulli modeling is expected to provide local word-order constraints as well as compensation of the monotonic decreasing against the word length in simple n-gram language model.

Futhermore, in practice, the frequency of a particular word-pair in a sentence is almost always either 0 or 1 (and $0! = 1! = 1$), neglecting the denominator of the weighting term leads the simple recursive calculation form for the Ngram-Bernoulli model, i.e.

$$\begin{aligned} \log \hat{P}_{\text{BGBER}}(\mathbf{W}_1^n) &= \log n + \log P(w_n|w_{n-1}) \\ &+ \log \hat{P}_{\text{BGBER}}(\mathbf{W}_1^{n-1}). \end{aligned} \qquad (3)$$

3.3. Relation with Insertion Penalty

As shown in Figure 1, negative WIP improves recognition accuracy in most cases and it can be related with (the enumerator $n!$ of the weighing term in) Ngram-Bernoulli modeling in the following way,

$$nQ = n \log q \leftrightarrow n \log n \approx \log(n!). \qquad (4)$$

Therefore, one can see that negative WIP is included in the Ngram-Bernoulli modeling, and no explicit WIP, needs to be incorpolated in Ngram-Bernoulli modeling.

4. EXPERIMENTAL EVALUATIONS

4.1. Experimental Setup

The effectiveness of Ngram-Bernoulli modeling is confirmed by comparing with a simple Ngram model and a normalized

Table 1: Analysis conditions

Sampling frequency	16 kHz
Quantization bit	16 bit
Window type	Hamming
Frame length	25 msec
Frame period	10 msec
Pre-emphasis	0.97
MFCC order	12
ΔMFCC order	12
Δpower order	1

version of the Ngram model. The normalized version of the Ngram probability is given by

$$P_{\text{NGNOR}}(\mathbf{W}) = \left\{ \prod_{i=1}^{n} P(w_i|\mathbf{w}_{i-N+1}^{i-1}) \right\}^{1/n}, \qquad (5)$$

where n is the length of word sequence.

Ngram probabilities are trained by ATR travel conversation corpora, containing 7,740 sentences with a vocabulary size of 4,784 words. All text is tagged using 27 POS categories. As the size of the corpora is not large, we used POS category based ngram for calculating the word-ngram. For example a word bigram is given by

$$P(w_i|w_{i-1}) \approx P(c_i|c_{i-1})P(w_i|c_i) \qquad (6)$$

where w_i falls into a POS category c_i.

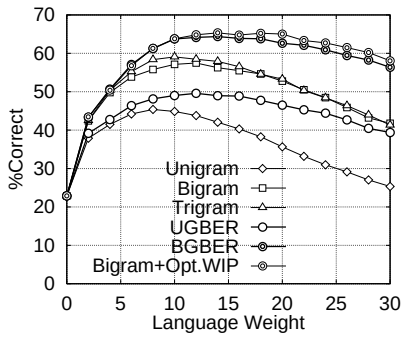
Triphone HMM's were used as the acoustic model. The HMM has 3 states and each state has 4 mixtures. Total number of states is approximately 12,000. We trained this model using 8,128 utterances from 54 speakers in the continuous speech corpus for research of the Acoustical society of Japan. Configurations of acoustic analysis conditions are listed in Table 1.

For the test utterance, one male speaker uttered a total of 150 sentences extracted from the training corpus. The test set consists of 5 subsets. Each subset contains 30 sentences of the same word-length, i.e. 9, 12, 15, 18, 21 words-long.

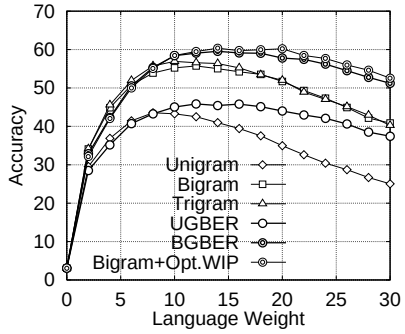
4.2. Results

Results of the experiments are given in Figure 2, 3 and 4. Throughout the figures, each model is referred as UGBER: Unigram-Bernoulli, BGBER: Bigram-Bernoulli, UGNOR: normalized version of Unigram, BGNOR: normalized version of Bigram.

Figure 2 shows that the performance of the Ngram-Bernoulli model is better than the simple Ngram at all values of language weight. The performance of the Bigram-Bernoulli model is better than simple trigram results. Furthermore, the



(a) Word Correct rate



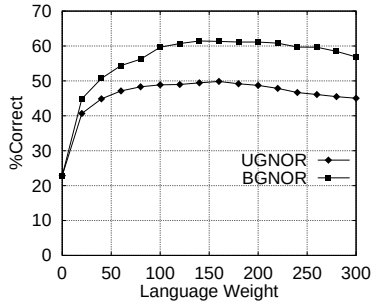
(b) Word Accuracy Score

Figure 2: Recognition performance across the language weight values. (N gram and N gram-Bernoulli models.)

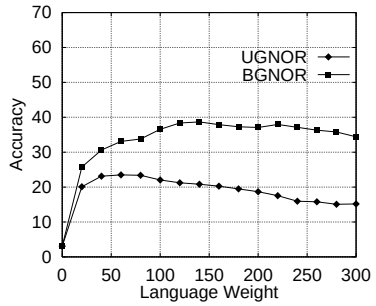
range of the language weight which gives the best recognition performance is wider in Ngram-Bernoulli than the simple Ngram model. With comparing to the case using WIP, the results of Ngram-Bernoulli model is comparable to the case when both language weight and WIP are carefully optimized by experiment.

The recognition performance of the normalized-ngram probability is shown in Figure 3. Note that the language weight value is different from the previous experiments. It can be seen that the word correct rate of the normalized ngram is comparable with the bigram-Bernoulli model. However, the word accuracy score is much lower than the Ngram-Bernoulli model. This result clarifies that the per-word evaluation of language probability causes frequent insertion errors.

The characteristics of each model can be discussed from the viewpoint of the average word length of recognition output, which is shown in Figure 3. The monotonic decreasing in average word length of the simple ngram model is not significant in both the Ngram-Bernoulli and the normalized Ngram. In the normalized Ngram, however, the averaged word length is longer than the real length (15 words in this experiment) due to frequent word insertion errors.

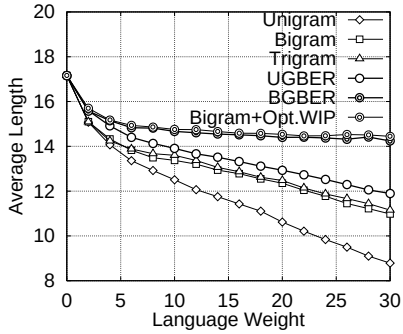


(a) Word Correct rate

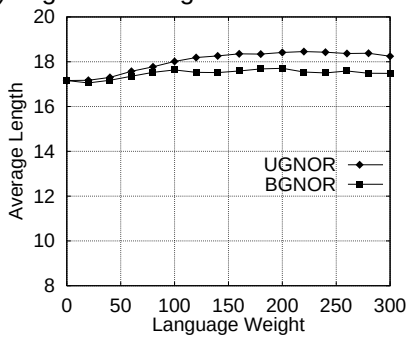


(b) Word Accuracy Score

Figure 3: Recognition performance across the language weight values. (Normalized N gram models.)



(a) N-gram and Ngram-Bernoulli Models



(b) Normalized Ngram Models

Figure 4: Average length of recognition result across the language weight values.

5. SUMMARY

In this paper we have discussed balancing acoustic and linguistic probabilities by extending the Ngram language modeling by taking the length of the word sequence into account. The basic framework of the modeling is to extend Bernoulli trial modeling of word occurrence to that of being conditioned by previous words. The effectiveness of the modeling is evaluated from the viewpoints concerned with the problems of simple Ngrams, 1) the balancing parameters are critical to recognition accuracy; 2) the optimal values of two parameters are related with each other; and 3) the optimal values are also governed by the length of word sequence. The experimental results show that the Ngram-Bernoulli modeling is effective for dealing with those problems.

6. REFERENCES

[1] L. R. Bahl, R. Bakis, F. Jelinek and R. L. Mercer, "Language-model/acoustic channel balance mechanism", *IBM Technical Disclosure Bulletin*, 23(7B), pp.3464-3465, Dec. 1980.

[2] F. Jelinek, gSelf-organized language modeling for speech recognition", in Readings in Speech Recognition, Morgan Kaufmann Publishers, Inc. 1990.

[3] A.J. Rubio et al., "On the influence of Frame-asynchronous grammar scoring in a csr system," *Proc. of ICASSP 97*, vol. 1, pp.895–898, April. 1997.

[4] S. J. Young, gThe HTK Hidden Morkov Model Toolkit: Design and Philosophy", CUED Technical Report F_INFENG/TR152, Cambridge University Engineering Department, Jul. 1994.