ESTIMATING ENTROPY OF A LANGUAGE FROM OPTIMAL WORD INSERTION PENALTY

Kazuya TAKEDA, Atsunori OGAWA and Fumitada ITAKURA
takeda@nuee.nagoya-u.ac.jp

Graduate School of Engineering
Nagoya University

ABSTRACT

The relationship between the optimal value of word insertion penalty and entropy of the language is discussed, based on the hypothesis that the optimal word insertion penalty compensates the probability given by a language model to the true probability. It is shown that the optimal word insertion penalty can be calculated as the difference between test set entropy of the given language model and true entropy of the given test set sentences. The correctness of the idea is confirmed through recognition experiment, where the entropy of the given set of sentences are estimated from two different language models and word insertion penalty optimized for each language model.

1. INTRODUCTION

One of the most important merits provided by stochastic framework in speech understanding systems, is the mathematical principal of combining two or more set of knowledge. Most notably, by utilizing linguistic knowledge in speech recognition through stochastic language modeling a very large vocabulary dictation system has been realized on a PC platform. The principal of combining acoustic an linguistic knowledge is given by Bay's rule i.e.

$$P(\mathbf{W}|\mathbf{A}) = \frac{P(\mathbf{A}|\mathbf{W})P(\mathbf{W})}{P(\mathbf{A})}. \qquad (1)$$

In the rule, disregarding denominator, a simple product of acoustic and linguistic probabilities gives a score to word sequence hypotheses even if the acoustic and linguistic models are estimated independently.

However, in the real system, it is widely known that balancing between acoustic and linguistic parameters is needed to optimize the system performance. The typical form of combining the two probabilities is

$$\log P(\mathbf{A}|\mathbf{W}) + \alpha \log P(\mathbf{W}) - nQ \qquad (2)$$

where α is known as *language weight* (LW), Q is known as *word insertion penalty* (WIP) and n is the number of words included in the sequence $\mathbf{W}$. Although utilizing the two parameters is quite common, very few descriptions have been given on the physical meanings and systematic optimization method of the parameters [1],[2]. Therefore, it is necessary to determine optimal value for each specific task, even if using the same acoustic and linguistic models.

The purpose of this paper is to relate optimal WIP and entropy of a language from a hypothesis that the optimal WIP can compensate the language probability given by a language model, i.e. word n-gram, to real probability. Furthermore, based on this hypothesis, a systematic way to determine the optimal WIP from the entropy of the test set is developed. The effectiveness of the method is evaluated by speech recognition experiments. The rest of the paper consists of the following sections. In the next section, optimality of WIP is discussed. In section 3, a method to estimate optimal WIP is proposed. In section 4, the method is improved by taking the word position within the sentence into account for the language modeling.

2. OPTIMALITY OF WORD INSERTION PENALTY AND ENTROPY

Although the scoring given in (2) is a common form of utilizing WIP and LW, we adopt the form below for the combination;

$$\log P(\mathbf{A}|\mathbf{W}) + \alpha\{\log P(\mathbf{W}) + nq\}. \qquad (3)$$

In this scoring, by enclosing the WIP in the parentheses, the sum, $\log P(\mathbf{W}) + nq$ is regarded as a new language probability.

From our preliminary experiments, the above formulation has proved to be effective for removing dependency between LW and WIP in optimization. Furthermore, it should be noted that, in (3), the WIP nq is added to the language score. We found in [4] that WIP should be used as a bonus to avoid shorter sentence preference in n-gram language modeling.

Therefore, introducing WIP is equivalent to compensating language model probability, $P_M(\mathbf{W})$, in log probability domain, by adding a score which is proportional to the number of words in the sentence;

$$\log \hat{P}_M(\mathbf{W}) = \log P_M(\mathbf{W}) + nq \qquad (4)$$

where $\hat{P}_M(\mathbf{W})$ is a compensated version of $P_M(\mathbf{W})$.

It is natural to think that the best recognition performance is given when the compensated version of language probability is equal to the true language probability. In that case, for the

optimal value of the WIP, q_{opt}, the below relation is expected to hold;

$$\log P(\mathbf{W}) = \log \hat{P}_M(\mathbf{W}) = \log P_M(\mathbf{W}) + nq_{opt} \quad . \quad (5)$$

The above relation can be rewritten as

$$q_{opt} = \frac{1}{n}\{-\ln P_M(\mathbf{W}) + \ln P(\mathbf{W})\} \quad . \quad (6)$$

Furthermore, calculating an optimal value of WIP, $\bar{q}_{opt}$, for a given set of test sentences,

$$\mathbf{S} = \{\mathbf{W}_k = w_1^{(k)} w_2^{(k)} \cdots w_{nk}^{(k)} | k = 1,2,\cdots,K\}$$

by averaging optimal WIPs calculated for each sentence, we can get,

$$\begin{aligned} \bar{q}_{opt} &= \frac{1}{K}\sum_{k=1}^{K} q_{opt}^{(k)} \\ &= \frac{1}{K}\sum_{k=1}^{K} \frac{1}{n_k}\{-\ln P_M(\mathbf{W}_k) + \ln P(\mathbf{W}_k)\} \end{aligned} \quad . \quad (7)$$

Assuming that the word probability is ergodic, finally, we can rewrite the optimal value of WIP in terms of the entropy of the language as follows.

$$\begin{aligned} \bar{q}_{opt} &= E\{q_{opt}^{(k)}\} \\ &= \frac{1}{\log_2 e} E\{-\ln P_M(\mathbf{W}_k) + \ln P(\mathbf{W}_k)\} \\ &= \frac{1}{\log_2 e}(H_M - H) \end{aligned} \quad (8)$$

where, H_M is the expected value of per-word log probability of the given language model calculated over test set $\mathbf{S}$, which is sometime referred as test set entropy, whereas H is the per-word entropy of the language $\mathbf{S}$.

3. ESTIMATING ENTROPY FROM OPTIMAL WIP

From the above discussions, WIP can be calculated as the difference between the entropy of the test set language and that of calculated by the given language model. Thus, if an optimal WIP is known for a particular language model, the true entropy of the test set, and the optimal WIP for different language models, can be estimated from (8). In this section, entropy of test set sentences is estimated from WIP's, each of them is optimized experimentally for different language models, to show the correctness of the above discussions.

3.1. Experimental Setup

As for language modes, word category base bigram and trigram are used, each of them are trained by ATR dialogue corpus [5]. The corpus consists of 7,740 sentences. The size of vocabulary is 4,784 and the number of word categories is 27. Trained word category n-grams are smoothed using Back-off Smoothing [6]. As for acoustic model, triphone HMM's consisting of 1000 states are used. Each state has four mixtures in the model. As for the feature parameters 12 MFCC and its delta are used with delta log power. For the test set, 150 sentences were extracted from the same corpus used for language model training. The 150 sentences were then read by a single male speaker and used for evaluation. For finding optimal values of LW and WIP, recognition experiments have been performed with variations in LW ranging from 2.0 to 30.0 in steps of 2.0 and WIP from –0.5 to 3.5 insteps of 0.5.

The entropy of the given language model, e.g. bigram, can be calculated by the following way.

$$H_{\text{bigram}} = -\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k - 1}\sum_{i=2}^{nk}\log_2\{P(h_i^{(k)}|h_{i-1}^{(k)})P(w_i^{(k)}|h_i^{(k)})\}$$

where $h_i^{(k)}$ is word category of $w_i^{(k)}$, $P(h_i^{(k)}|h_{i-1}^{(k)})$ is word category bigram probability and $P(w_i^{(k)}|h_i^{(k)})$ is word probability of $w_i^{(k)}$ within the word class $h_i^{(k)}$.

The test set entropy of each language model is 6.38 [bit] for word category bigram and 6.13 [bit] for word category trigram, respectively.

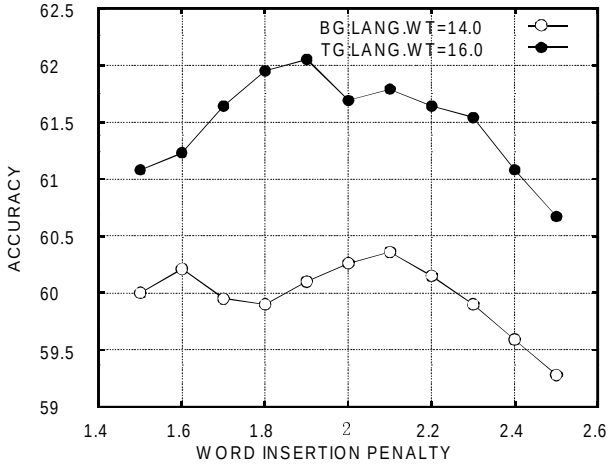

Figure 1: Relationship between WIP and word accuracy at the optimal LW values. BG and TG shows the results of word category bigram and trigram respectively.

3.2. Experimental Results

The recognition accuracy at various WIP values are shown in Figure 1. From the figure, it can be seen that optimal value of WIP is 2.1 for word category bigram and 1.9 for trigram respectively. True per-word entropy of the test set calculated from each results is;

- word category bigram case
  $H \approx 6.38 - \log_2 e \times 2.1 \approx 3.35$ [bit]
- word category trigram case
  $H \approx 6.13 - \log_2 e \times 1.9 \approx 3.39$ [bit]

Thus, it is confirmed that almost the same entropy is estimated from different combinations of language model and WIP.

The perplexity calculated from the entropy values are 10.2 and 10.5, respectively. These values are reasonable as the average branching factor of a language consists of 150 sentences.

4. SETTING WIP DEPENDING ON THE SENTENCIAL LOCATION

4.1. Extending WIP

In the previous sections, WIP is modeled as a constant thoroughout the sentence. This is based on the assumption that the word probability is independent from the location in the sentence. In this section, WIP is reformulated to be depend upon the sentential location.

First, extending the basic WIP form of (4) to

$$\ln P(\mathbf{W}) = \ln P_M(\mathbf{W}) + \sum_{i=1}^{n} q(i) \qquad (9)$$

where, n is the number of words included in the word sequence $\mathbf{W}$. Note that WIP is not value but function of the sentential location.

The optimal WIP function is, then, also extended from (6) to the form of,

$$q_{\text{opt}}(i) = -\ln P_M(w_i|w_{i-1}) + \ln P(w_i|w_1 w_2 \cdots w_{i-1}), \quad (10)$$

for the bigram case, for example. As in the previous section, we can calculate the optimal WIP function, $\bar{q}_{\text{opt}}(i)$, for a given set of sentences $\mathbf{S}$. Here, $\bar{q}_{\text{opt}}(i)$ is given as the average of $q_{\text{opt}}(i)$ over the set of word sequences, $\mathbf{W}_k$, which satisfies $n_k \geq i$, as follows.

$$\begin{aligned}\bar{q}_{\text{opt}}(i) &= \frac{1}{M(i)} \sum_{\substack{k=1 \\ (nk \geq i)}}^{K} q_{\text{opt}}(i) \\ &= \frac{1}{M(i)} \sum_{\substack{k=1 \\ (nk \geq i)}}^{K} \left[\begin{array}{l} -\ln P_M(w_i^{(k)}|w_{i-1}^{(k)}) \\ \quad + \ln P(w_i^{(k)}|w_1^{(k)} w_2^{(k)} \cdots w_{i-1}^{(k)}) \end{array} \right] \qquad (11) \\ &= \frac{1}{\log_2 e} \left\{ \begin{array}{l} H_M(W_i^{(k)}|W_{i-1}^{(k)}) \\ \quad - H(W_i^{(k)}|W_1^{(k)} W_2^{(k)} \cdots W_{i-1}^{(k)}) \end{array} \right\}\end{aligned}$$

where $M(i)$ is the number of the sentences which contains more than n_k words, and W_i is a random variable corresponding to the word w_i. $H_M(W_i|W_{i-1})$ is a conditional entropy of word W_i preceded by the word W_{i-1}, whereas $H(W_i|W_1 W_2 \cdots W_{i-1})$ is a conditional entropy of word W_i following the given word sequence $W_1 W_2 \cdots W_{i-1}$. In n-gram modeling, which assumes the word sequence as a Markov process, $H_M(W_i|W_{i-1})$ is constant to i. On the other hand, $H(W_i|W_1 W_2 \cdots W_{i-1})$ is monotonically decreasing function of i. Thus, it is expected to converge to the true entropy of the language H. Furthermore, in general,

$$H_M(W_i|W_{i-1}) \geq H(W_i|W_1 W_2 \cdots W_{i-1}) \quad (12)$$

holds.

From above discussions, $\bar{q}_{\text{opt}}(i)$ is expected to increase monotonically as i increase and converge to a finite value, as shown in Figure 2. Thus, let γ_{opt} the limit value of $\bar{q}_{\text{opt}}(i)$ of $i \to \infty$, per-word entropy of the language $\mathbf{S}$ is calculated by

$$\begin{aligned}H &= \lim_{i \to \infty} H(W_i|W_1 W_2 \cdots W_{i-1}) \\ &= \lim_{i \to \infty} \{H_M(W_i|W_{i-1}) - \log_2 \bar{q}_{\text{opt}}(i)\} \qquad (13) \\ &= H_M(W_i|W_{i-1}) - \log_2 \bar{\gamma}_{\text{opt}}\end{aligned}$$

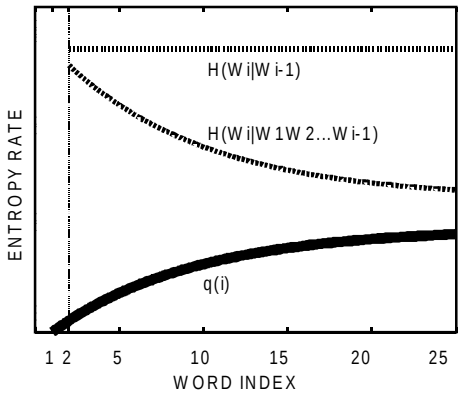

Figure 2: Global shape of the entropy of bigram language model $H_M(W_i|W_{i-1})$, true conditional entropy $H(W_i|W_1 W_2 \cdots W_{i-1})$ and the insertion penalty function $q_{\text{opt}}(i)$.

4.2. Recognition Experiments

In order to confirm the effectiveness of changing WIP depending on the sentential location, recognition experiments to estimate the per-word entropy H have been performed on the same test set $\mathbf{S}$ as the previous section. In the experiments, $q(i)$ is parameterized by the below form and recognition performance is measured for various combinations of γ and β.

$$q(i) = \gamma(1 - \beta^{1-i}). \qquad (14)$$

This form is based on our previous results that the recognition accuracy was improved when $q(i) = \ln(i)$ [4]. The above function approximates the logarithmic function where $10 \leq i \leq 25$ and converges to γ when $i \to \infty$ as shown in Figure 3.

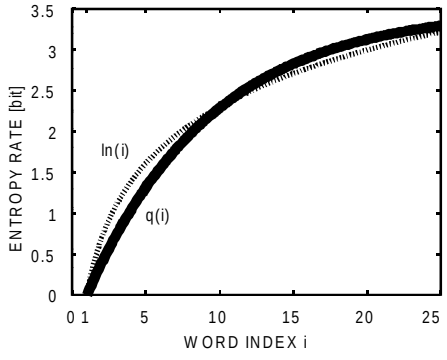

Figure 3: An example of $q(i)$ where $\beta = 1.125$ and $\gamma = 3.5$.

The obtained results are shown in Figure 4. From the figure, the optimal value of γ is approximated 2.0. This is a similar value with that of using fixed WIP value in the previous section. Since the obtained recognition accuracy is slightly better than the fixed WIP method, the effectiveness of using sentential location dependent WIP is confirmed.

The estimated per-word entropy is calculated as

$$H = H_M(W_i|W_{i-1}) - \log_2 e \cdot \gamma_{\text{opt}} \approx 6.04 - \log_2 e \cdot 2.0 \approx 3.16 \ [\text{bit}], \quad (15)$$

and the estimated perplexity value 8.9 is lower than the estimation of the previous section.

The reason for this inconsistency is the difference between conditional entropy $H_M(W_i|W_{i-1})$ in (15) and test set entropy of the bigram H_M in (8). It should be noted that the two entropy values are calculated in different ways as shown in below.

- Test set entropy is calculated by assuming that the conditional probability is independent from sentential locaion.

$$-\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k - 1}\sum_{i=2}^{nk}\log_2 P(w_i^{(k)}|w_{i-1}^{(k)})$$

- Conditional entropy is calculated by assuming that the conditional probability is depending on the sentential location.

$$-\frac{1}{N-1}\sum_{i=2}^{N}\frac{1}{M(i)}\sum_{\substack{k=1 \\ (nk \geq i)}}^{K}\log_2 P(w_i^{(k)}|w_{i-1}^{(k)}) .$$

5. SUMMARY

In this paper, we have discussed the method of determining optimal WIP from the viewpoint of entropy of a language. Based on the hypothesis that the WIP compensates the language probability given by a model to a real probability, optimal WIP is formulated as a difference between per-word test set entropy and true entropy of test set. The recognition experiments have reveild that when estimating true entropy by optimal WIP, the same entropy is estimated in both bigram and trigram language models. Therefore, the relation between WIP and entropy of a language is confirmed.

In this paper, sentential location dependent WIP is also formulated. Although the converged value of WIP function is very similar to the fixed WIP value, estimated entropy does not coincide with the value estimated from fixed WIP. This discrepancy results from using different methods to calculate model entropy.

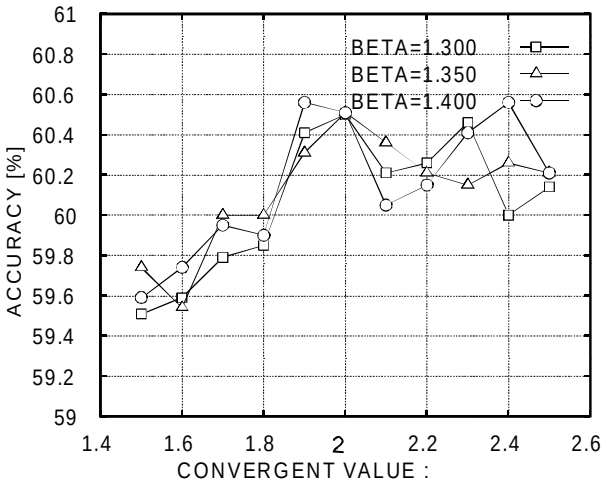

Figure 4: Recognition accuracy at various combination of parameters to represent $q(i)$. For this experiment LW α is fixed to be 14.0.

6. REFERENCES

1. F.Jelinek, "Continuous speech recognition by statistical methods," Proc. IEEE, Vol.64, No.4, pp.532-556, Apr. !976
2. L.R.Bahl et al. "Language-model / acoustic channel balance mechanism", IBM Technical Disclosure Bull. 23 (7B), pp.3464-3465, Dec. 1980
3. A.J.Rubio et al., "On the influence of frame-asynchronous grammar scoring in a CSR system," Proc. of ICASSP 97, vol.II, pp.895-898, April. 1997.
4. A. Ogawa et al., "Language Modeling for Robust Balancing of Acoustic and Linguistic Probabilities," Proc. of ICASSP 98, vol.I, pp.181-184, May, 1998..
5. T.Ehara et al., "Contents of the ATR Dialogue Database," ATR Technical Report, Oct. 1990, (in Japanese)
6. S.M.Katz, "Estimation of probabilities from sparse data for language model conponent of a speech recognizer," IEEE Trans. ASSP-35, No.3, pp.400-401