# T2R2 東京科学大学 リサーチリポジトリ
## Science Tokyo Research Repository

/

## Article / Book Information

| | |
|---|---|
| Title | Sentence selection by direct likelihood maximization for language model adaptation |
| Authors | Takahiro Shinozaki, Yu Kubota, Sadaoki Furui, Eiji Utsunomiya, Yasutaka Shindoh |
| Citation | Proc. INTERSPEECH 2011, , , pp. 613-616, |
| Pub. date | 2011, 8 |
| Copyright | (c) 2011 International Speech Communication Association, ISCA |
| DOI | http://dx.doi.org/ |
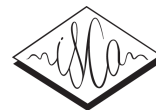
# Sentence Selection by Direct Likelihood Maximization for Language Model Adaptation

Takahiro Shinozaki[1,3], Yu Kubota[1], Sadaoki Furui[1], Eiji Utsunomiya[2], Yasutaka Shindoh[2]

[1]Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan
[2]Technology Development Center, KDDI R&D Labs. Inc., Tokyo, Japan
[3]Division of Information Sciences, Chiba University, Chiba, Japan

## Abstract

A general framework of language model task adaptation is to select documents in a large training set based on a language model estimated on a development data. However, this strategy has a deficiency that the selected documents are biased to the most frequent patterns in the development data. To address this problem, a new task adaptation method is proposed that selects documents in the training set so as to directly reduce the perplexity on the development set. Moreover, a weighting method to modify the perplexity objective function is proposed to improve the generalization to unseen data. The proposed adaptation methods are evaluated by large vocabulary speech recognition experiments. It is shown that the proposed adaptation with the weighting term produces a compact-size model that gives consistently lower word error rates for different tasks.

**Index Terms**: language model, selective training, task adaptation, N-gram model, N-gram count

## 1. Introduction

Performance of an N-gram language model largely depends on the matching of training and test conditions. Therefore, task adaptation is important. Given a large task-independent training data and a small task-dependent development data, a general framework of language model adaptation is to first estimate a model using the development data. Then, probabilities of observing each document in the training set are evaluated using the model. A subset of the training set is formed by gathering documents with higher probabilities, which is assumed to better match the target task. Finally, a task adapted model is made using that subset [1].

However, there is no guarantee with this procedure that the adapted model gives lower development-set perplexity. This is because the direction of model training and evaluation is opposite in the document selection and the final model training. In fact, this procedure tends to select documents that match the most frequent patterns in the development-set. As the result, less frequent patterns in the development set are likely to be ignored and the development-set perplexity increases.

A solution to this problem is to train models for all possible subsets of the training set and pick a model that gives the lowest perplexity for the development set. This idea is simple and direct, but a problem exists in high computational cost. The number of possible subsets of the training set is $2^K$ where $K$ is the number of documents. For a large database, $K$ can be on the order of millions. Therefore, it is impossible to try all the subsets.

In order to implement the idea with feasible computational cost, we propose Direct Likelihood Maximization Selection (DLMS) method that introduces a greedy search approximation and an efficient language model probability evaluation algorithm based on N-gram counts. Moreover, we propose Context Locality Weight (CLW) that is used to improve generalization of DLMS for N-grams that do not appear in the development set.

The framework of DLMS is similar to the selective training proposed for acoustic model [2]. The contribution of this study is to create the language model version of the selective training and to introduce the CLW to the selection process. We apply the proposed methods to a task using a very large Blog database as the training set and a relatively small amount of task dependent transcriptions as the development set. Large vocabulary speech recognition is performed and the model performance is evaluated by word error rates.

The organization of this paper is as follows. In Section 2, the conventional and proposed language model task adaptation methods are described. Experimental conditions are described in Section 3 and the results are shown in Section 4. Conclusions and future works are given in Section 5.

## 2. Language model adaptation based on data selection

In this section, first a conventional language model task adaptation method is briefly reviewed, which is referred to as indirect selection method in this study and is used as a baseline. Then, proposed Direct Likelihood Maximization Selection (DLMS) and Context Locality Weight
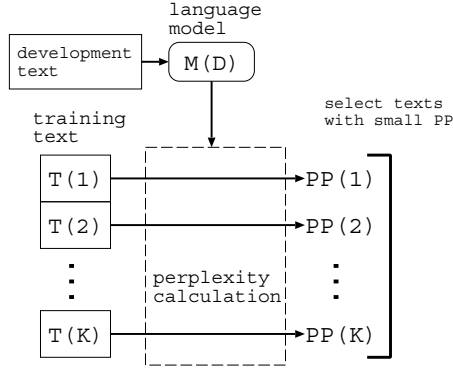
28 − 31 August 2011, Florence, Italy

Figure 1: Framework of the indirect selection method.



Figure 2: Framework of the DLMS method.

(CLW) are explained. Both of the selection methods assume a large task-independent training set and a small task-dependent development set.

## 2.1. Indirect selection

Figure 1 shows the framework of the Indirect selection method [1]. In this procedure, first a reference language model $M(D)$ is trained on a development set. Then, perplexity $PP(k)$ of each document $T(k)$ in the training set is evaluated using that model. With a proper threshold $\theta$, documents that have smaller perplexity than $\theta$ are selected. A task adapted model is trained using the selected set of documents as training data.

A problem of this method is that the selected documents are biased to the most frequent N-grams in the development set. To illustrate the problem, let a development set contains two kinds of uni-grams "a" and "b", and let their occurrences are 7 and 3, respectively. Let a training set contains two documents. The first document has 7 occurrences of uni-gram "a" and 3 occurrences of "b". Similarly, let the occurrences of the uni-gram "a" in the second document is 9 and "b" is 1. With this setting, the probability of the first document is $\left(\frac{7}{10}\right)^7 \cdot \left(\frac{3}{10}\right)^3$, and the second one is $\left(\frac{7}{10}\right)^9 \cdot \left(\frac{3}{10}\right)^1$. While the uni-gram distribution of the first document best matches the development data, the highest probability or the lowest perplexity is given by the second document that has the highest occurrence of uni-gram "a".

## 2.2. Direct Likelihood Maximization Selection (DLMS)

The problem of the indirect selection method is due to the fact that the direction of model training and evaluation is opposite in the document selection and the final model training. If the development-set perplexity is directly used as the objective score of the document selection, this problem is solved. However, possible number of subsets of a training set is exponential to the number of documents included in the training set, and it is not feasible to try them all. Therefore, we adopt a greedy approximation as shown in Figure 2.
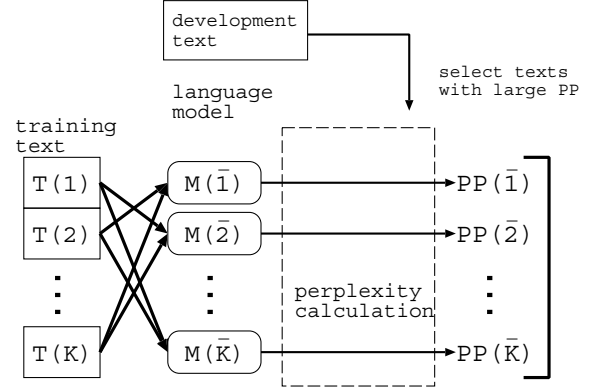
In the proposed DLMS method, for each document $T(k)$ in a training set, a language model $M\left(\bar{k}\right)$ is trained using the training set after removing document $T(k)$, and perplexity $PP\left(\bar{k}\right)$ of a development set is evaluated using that model. Lower perplexity indicates it is better to remove the corresponding documents. With a proper threshold $\theta$, all documents having the perplexity larger than $\theta$ are selected simultaneously. By ignoring combinatorial effect of document selection, the number of models to be evaluated is reduced from exponential to linear to the number of documents $K$. A task adapted model is trained using the selected set of documents as training data.

While the computational cost is largely reduced by the approximation that ignore the combinatorial effect, it still requires significant amount of computation if a language model is made from scratch for each document $T(k)$ to be removed. To further reduce the computational cost, we utilize N-gram counts as shown in Equation (1), to evaluate development set probability $P_{(\bar{k})}(w_1^n)$ after removing document $T(k)$.

$$P_{(\bar{k})}(w_1^n) = \prod_{j=1}^{n} \frac{c(w_{j-N+1}^j) - c_k(w_{j-N+1}^j)}{c(w_{j-N+1}^{j-1}) - c_k(w_{j-N+1}^{j-1})} \quad (1)$$

In the equation, $w_1^n$ indicates a $n$-length word sequence, $c(w_i^j)$ is counts of $(j - i + 1)$-length N-gram $w_i^j$ in the training set, $c_k(w_i^j)$ is counts of N-gram $w_i^j$ in $k$-th document $T(k)$. Thus, by subtracting counts in $T(k)$ from the counts in entire training set, N-gram probability after removing $T(k)$ is efficiently obtained.

When an N-gram that does not appear in the training subset occurs in the development set, $(N-1)$-gram probability is used instead. However, back-off weights (e.g. [3]) are not considered as it increases computational cost.

The "document" used as the unit for the selection can be a sentence or a set of sentences. The smaller the unit, the finer the selection becomes, but with the trade-off of the computational cost.

### 2.3. Context locality weight (CLW)

The proposed DLMS directly selects a training subset that maximizes the development set perplexity. However, in terms of generalization to unseen evaluation data, it may be useful to introduce a modification to the perplexity objective score. A problem of using the perplexity score is that it gives the highest score when all the N-grams that do not appear in the development set are removed from the training set. Apparently, this is not suitable since the development data is usually limited and there are many unseen but important N-grams for its application.

To address this problem, we propose and introduce CLW for each word probability to estimate the perplexity objective function as shown in Equation (2).

$$\hat{P}_{(\bar{k})}(w_j|w_{j-N+1}^{j-1})$$
$$= P_{(\bar{k})}(w_j|w_{j-N+1}^{j-1})\left(1 - \frac{c_k(w_{j-N+1}^{j-1})}{c(w_{j-N+1}^{j-1})}\right) \quad (2)$$

In the equation, $P_{(\bar{k})}(w_j|w_{j-N+1}^{j-1})$ is N-gram probability estimated on the training set after removing $k$-th document $T(k)$, $\left(1 - \frac{c_k(w_{j-N+1}^{j-1})}{c(w_{j-N+1}^{j-1})}\right)$ is the proposed weight term, and $\hat{P}_{(\bar{k})}(w_j|w_{j-N+1}^{j-1})$ is the N-gram probability after the compensation. The meaning of the weight is that if an N-gram context $w_{j-N+1}^{j-1}$ appears specifically in a document $T(k)$, the ratio $\frac{c_k(w_{j-N+1}^{j-1})}{c(w_{j-N+1}^{j-1})}$ approaches to 1.0 and CLW becomes close to 0.0. Thus, CLW can be regarded as a kind of document frequency of a N-gram context. A document that includes N-grams with low CLW is unique compared to others, and it worth for an attention from the document selection point of view. If such document share the N-gram contexts with the development set, it would be useful to select it for the training subset. By incorporating CLW to the objective function as shown in Equation (2), a document has larger chance to be selected if it include unique N-gram contexts that appears in the development set.

## 3. Experimental setups

The proposed language model adaptation methods are evaluated by large vocabulary speech recognition experiments. A Blog database is used as the training set. It includes 555 million words and 25 million sentences. The data is split into 10 sentences chunks and they were used as the document unit.

Two recognition tasks are used. One is a simulated presentation task and the other is an academic presentation task. Both of them are official test sets of Corpus of Spontaneous Japanese (CSJ) [4]. The simulated presentation task consists of 10 presentation given by 10 different speakers including males and females. The academic presentation task consists of 10 academic presentations given by 10 different male speakers. The simulated presentations are about everyday topics and they are closer to the Blog data than the academic presentations. However, since Blog is a written text, it is still different from the spoken simulated presentations. A development set for the simulated presentation task is a set of simulated presentations from CSJ. Similarly, a development set for the academic presentation task is a set of academic presentations. The length of each presentation in the development and test sets is around 10 minutes.

All language models are a trigram with 30k vocabulary. N-grams that occur less than three times are cut-off. Acoustic model is a triphone HMM having 3000 states in total. Each HMM state has 32 Gaussian components. The parameters are estimated by ML and MPE training [5] using 232 hours of CSJ academic presentations. MFCC based acoustic features were used with 39 elements comprising 12 MFCCs, log energy, their deltas, and delta-deltas. Speech recognition is performed using the $T^3$ decoder [6].

## 4. Experimental results

Figure 3 shows the ratio of selected training subset and averaged word error rates for the simulated presentation task. The development set consisted of 100 simulated presentations and HMM is estimated by ML training. In the Figure, "Random" is a result by random selection, "Baseline" is the conventional indirect selection method, "DLMS" is the proposed DLMS method, and "DLMS-CLW" is DLMS with CLW compensation. The subset ratio 1.0 means the model is estimated using all the original training set, which is actually a task independent model. As can be seen, the lowest word error rates by the baseline selection method and the DLMS are similar. However, the DLMS achieved the minimum word error rate with smaller training subset size than the baseline. This results in smaller model size as shown in Table 1 and contributes to reduce memory size in decoding. The baseline method gave worse results than the random selection when ratio of selected documents were less than 0.05. This is due to the bias effect described in sub-section 2.1. When the CLW is introduced to DLMS, word error rates were further reduced. Relative word error rate reduction by the baseline, DLMS, and DLMS with CLW methods from the task independent model were 1.2%, 1.4%, and 3.1%, respectively. The difference between the baseline and the DLMS with CLW was statistically significant. CPU time to execute DLMS was about 250 hours.

Table 2 shows the relationship between the size of development data and word error rates. The result by the random selection is independent of the development set size. While the baseline and the DLMS gave improvement when more than 25 presentations were used as the development data, DLMS with CLW improved the recog-
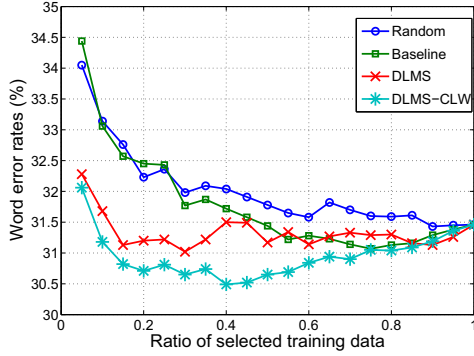
Figure 3: Amount of selected training subset and averaged word error rates for the simulated presentation task.

Table 1: Number of 3-grams in a language model that gave minimum word error rate

|          | # of 3grams |
|----------|-------------|
| Random   | 1.5E7       |
| Baseline | 1.2E7       |
| DLMS     | 5.6E6       |
| DLMS-CLW | 9.4E6       |

nition performance even when only 5 presentations were used.

As a supplemental experiment, word error rates were evaluated when the development data were mixed to the selected documents to train adapted models. The error rates slightly reduced but the tendencies were the same. When 100 presentations were used as the development data with this condition, the error rates by the baseline, the DLMS, and the DLMS-CLW were 30.9, 31.0, and 30.4, respectively, whereas it was 31.5% when the task independent model was used.

As mentioned in the experimental setup, the topics of the simulated presentations are more or less similar to the ones found in the Blog data. To see the effect of the proposed adaptation method when the topics are more different, the academic presentation task was used as the target and the 25 academic presentations were used as its development set. Figure 4 shows the results using the MPE-trained acoustic model. As can be seen in the figure, the proposed DLMS gave improvement over the task independent model while almost no improvement was observed by the baseline method. Relative word error rate reduction by the baseline, the DLMS and the DLMS with

Table 2: Size of development data and word error rates

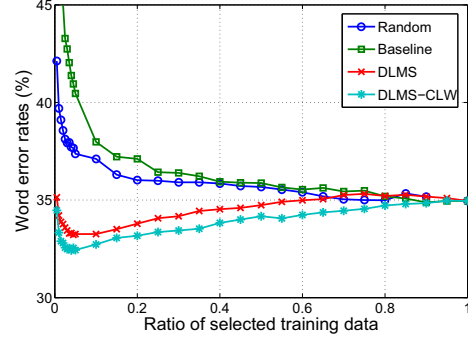| # of presentations | 5 | 10 | 25 | 100 |
|---------------------|------|------|------|------|
| Random   | 31.4 | 31.4 | 31.4 | 31.4 |
| Baseline | 31.5 | 31.4 | 31.3 | 31.1 |
| DLMS     | 31.5 | 31.5 | 31.3 | 31.0 |
| DLMS-CLW | 31.2 | 30.9 | 30.7 | 30.5 |



Figure 4: Amount of selected training subset and averaged word error rates for the academic presentation task.

CLW were 0.2%, 5.0% and 7.3%, respectivelly. The differences between these results were all statistically significant.

## 5. Conclusion

The DLMS language model task adaptation method has been proposed that is based on selecting a subset of a training set so as to directly maximize development set perplexity. Speech recognition results show that the DLMS gives similar or lower word error rates than the conventional indirect selection method, and it is effective to obtain a small-size model keeping the recognition performance. Moreover, when the proposed weighting term CLW was introduced to the DLMS, it gives consistently lower word error rates than the indirect selection method. Future work includes comparisons with other methods such as the relative entropy based sentence selection [7]. It would be interesting to extending the DLMS for unsupervised adaptation.

## 6. References

[1] R. Nisimura, K. Komatsu, Y. Kuroda, K. Nagatomo, A. Lee, H. Saruwatari, and K. Shikano, "Automatic n-gram language model creation from web resources," in *Proc. Eurospeech*, 2001, pp. 2127–2130.

[2] T. Cincarek, T. Tomoki, H. Saruwatari, and K. Shikano, "Utterance-based selective training for the automatic creation of task-dependent acoustic models," *IEICE Transactions on Information and Systems*, vol. E89-D, no. 3, pp. 962–969, 2006.

[3] S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recogniser," *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 35, no. 3, pp. 400–401,, 1987.

[4] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui, "Benchmark test for speech recognition using the Corpus of Spontaneous Japanese," in *Proc. SSPR2003*, 2003, pp. 135–138.

[5] D. Povey and P. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, vol. I, 2002, pp. 105–108.

[6] P. R. Dixon, D. A. Caseiro, T. Oonishi, and S. Furui, "The titech large vocabulary wfst speech recognition system," in *Proc. IEEE ASRU*, 2007, pp. 443–448.

[7] A. Sethy, P. Georgiou, and S. Narayanan, "Text data acquisition for domain-specific language models," in *Proc. EMNLP*, 2006, pp. 382–389.