LETTER

N-gram Adaptation with Dynamic Interpolation Coefficient Using Information Retrieval Technique

Joon-Ki CHOI[†a)], ***Student Member and*** **Yung-Hwan OH**[†], ***Nonmember***

SUMMARY This study presents an N-gram adaptation technique when additional text data for the adaptation do not exist. We use a language modeling approach to the information retrieval (IR) technique to collect the appropriate adaptation corpus from baseline text data. We propose to use a dynamic interpolation coefficient to merge the N-gram, where the interpolation coefficient is estimated from the word hypotheses obtained by segmenting the input speech. Experimental results show that the proposed adapted N-gram always has better performance than the background N-gram.

key words: *language model adaptation, adaptation corpus, dynamic interpolation coefficient, speech recognition*

1. Introduction

N-gram adaptation has gained popularity due to its ability to cope with problem of domain dependency in N-gram language model [1]. The N-gram adaptation technique updates the characteristics of the background N-gram model into a domain-specific model with little or no manually annotated adaptation corpus. The two major problems which determine the performance of the final adapted N-gram are the acquisition of the adaptation corpus and the combining method of the background and the adapted model. Firstly, we propose the usage of bigram and trigram retrieval model in the language modeling approach to information retrieval (IR) to collect the adaptation corpus. Recently, IR techniques have been widely used to build a training corpus for an N-gram adaptation [2], [3]. Among the various IR techniques, the language modeling approach to IR can directly measure similarity between the language model of a query and the language model of a target document [4]. Experimental results show that the usage of bigram and trigram retrieval model, instead of unigram model which is popular in text based IR, improves the quality of the collected adaptation corpus.

Secondly, we propose to use a dynamic language model interpolation coefficient to solve the problem of the language model merging. The proposed interpolation coefficient varies according to the segment of the recognition hypothesis. Previously, a global optimal static interpolation coefficient [2] and a history-dependent coefficient [5] were determined by EM algorithm with held-out validation data. However, it is erroneous that the short history of the held-out validation data should decide the interpolation coefficient, as the validation data cannot represent precisely the structure of the test data. Kalai proposed an on-line algorithm to estimate the dynamic interpolation coefficient using the history of each word hypothesis without any validation data [6]. However, due to a recognition error, the optimization for every path of the sentence hypothesis can be wrong. In our proposed method, all word hypotheses in a certain segment of the input speech determine one interpolation coefficient for the segment. This would alleviate the effect of inappropriate validation data and the recognition error.

This study is organized as follows. The following Sect. 2 discusses the acquisition of the adaptation corpus. Section 3 describes how to estimate the dynamic interpolation coefficient with the recognition hypothesis. This is followed by the experimental results and finally by the conclusion.

2. Collecting Adaptation Corpus Using IR

The adaptation of the N-gram probability with same vocabulary is focused on in this study. Hence the adaptation corpus is retrieved from only the baseline text corpus. In addition, the lexicon adaptation technique is not used.

While the adaptation corpus is collected by IR technique, recognition hypotheses such as a word hypothesis or a sentence hypothesis are used as a query in IR. Hence the errors of speech recognition can degrade the quality of the retrieved adaptation corpus. To reduce the effect of the recognition error, a word level confidence measure is used as a query selection criterion. We use the link posterior probability from Mangu's confusion network as the confidence measure [7]. The word hypothesis that has a lower confidence measure than a certain threshold is regarded as a recognition error and hence discarded.

From among various IR techniques, we used the language modeling approach to IR to acquire the appropriate adaptation corpus. In this method, the retrieved documents are ranked based on the probability of producing a query from the corresponding language models of these documents [4]. For this reason, the language modeling approach, as opposed to other IR methods, can find the documents that have a similar N-gram distribution to the input speech. The Kullback-Leibler divergence score is used as the distance measurement between document d and query q, as described by following equation:

Manuscript received December 28, 2005.
Manuscript revised March 29, 2006.

[†]The authors are with the Division of Computer Science, KAIST, Daejeon, Korea.
a) E-mail: jkchoi@speech.kaist.ac.kr

DOI: 10.1093/ietisy/e89–d.9.2579

$$Div(M_q \parallel M_d) = \sum_w p(w|M_q) \log \frac{p(w|M_q)}{p(w|M_d)} \quad (1)$$

where M_q denotes the language model of query q, and M_d denotes the language model of document d. The concept of the ordinary language model is applied to $p(w|M_d)$ as follows:

$$p(w|M_d) = \begin{cases} p_s(w|M_d) & w \in d \\ \alpha_d p(w|M_c) & \text{otherwise} \end{cases} \quad (2)$$

where $p_s(w|M_d)$ is the probability that the word w is seen in document d, α is a document-dependent constant which controls the probability of unseen words, and M_c is the language model of the full-text corpus. Given Eq. (2), the right side of the Eq. (1) can be represented as Eq. (3), where the detailed derivation can be found in [9].

$$\sum_{w \in d \cap q} p(w|M_q) \log \frac{p_s(w|M_d)}{\alpha_d p(w|M_c)} + \log \alpha_d \quad (3)$$

There are many language model smoothing techniques to evaluate $p(w|M_d)$ and α_d. Previously, only unigram retrieval model was used to collect the adaptation corpus [2]. However, since our ultimate goal is to retrieve documents which have the closest N-gram probabilities to those of the incoming test speech sentence, we used Miller's bigram and trigram retrieval model [8] to collect the adaptation corpus. In the case of a bigram, Eq. (3) can be rewritten as follows:

$$\sum_{w_i w_{i-1}} p(w_i|M_q, w_{i-1}) \log \frac{p_s(w_i|M_d, w_{i-1})}{\alpha_d p(w_i|M_c, w_{i-1})} + \log \alpha_d \quad (4)$$

where $w_i w_{i-1}$ is the common bigram entry in document d and query q. Similarly, this formula can be expanded to a trigram model case.

3. Dynamic Interpolation Coefficient

The raw adapted N-gram, which is built with only an adaptation corpus, suffers from a data sparseness problem. To cope with this problem, we used simple and effective linear interpolation to build the merged model $\hat{p}$. The equation is as follows:

$$\hat{p}(w|h) = \lambda p_a(w|h) + (1 - \lambda) p_b(w|h) \quad (5)$$

where p_a is the raw adapted N-gram, p_b is the background N-gram, h is word history of word w, and λ is the interpolation coefficient that adjusts the contribution of each component model. Notice that if some parts of the input speech are better represented by the raw adapted N-gram than the background N-gram, then those parts of the speech should be rescored by a merged model that has a larger weight λ on the raw adapted model p_a than on the background model p_b. Therefore, we segment the input test speech and give different interpolation coefficient λ to different segment. Compared to the conventional method where λ of a word is determined by only the preceding history of the word, our proposed method estimates λ using all word hypotheses which belong to the same segment.

In order to obtain appropriate λ, the incoming speech must be segmented effectively. For adequate segmentation, we exploit the sensitivity of the N-best list to a change of the language model value. The segmenting procedure is as follows:

1. **Rescoring**
   - Rescore the lattice with the raw adapted language model and extract the N-best list N_a
   - Rescore the lattice with the background language model and extract the N-best list N_b
2. **Aligning**
   - Align the N-best list sets with a Levenshtein distance. Let $W_i(a)$, $W_i(b)$ denote the set of word hypotheses at the same position (ith) in N_a, and N_b respectively
3. **Grouping**
   - Group unmatched word hypotheses as a segment which satisfy consecutive interval (i, j) that for all $i \le k \le j$, $W_k(a) \neq W_k(b)$
   - Group remaining each consecutive word hypotheses as a segment respectively

An example of the segmenting procedure is depicted in Fig. 1 where the N-best list is represented as a lattice for easy understanding.

Using the word hypotheses in the segment as held-out validation data, it is possible to find an interpolation coefficient that is biased toward the correct language model by an EM algorithm [10]. The update step from λ_n to λ_{n+1} is expressed by Eq. (6):

$$\lambda_{n+1} = \frac{1}{M} \sum_{m=0}^{M-1} \frac{\lambda_n p_a(w_m|h_m)}{\lambda_n p_a(w_m|h_m) + (1 - \lambda_n) p_b(w_m|h_m)} \quad (6)$$

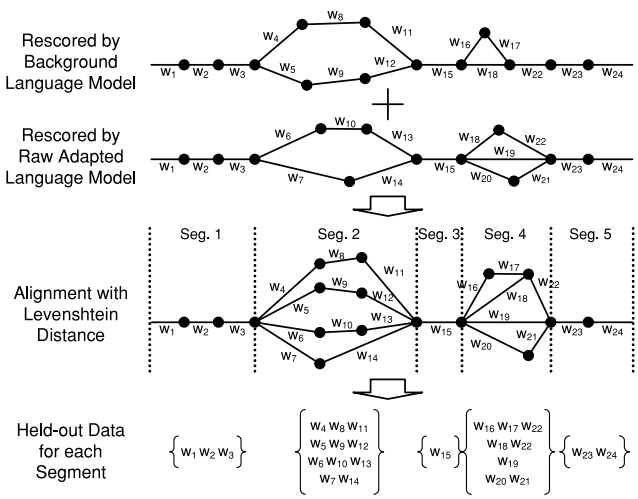

Fig. 1 Speech segmentation with sensitivity of the N-best list to the change of the language model value. Align the rescored N-best lists using two language models. After segmenting with this alignment, the held-out validation data is extracted from corresponding segment.

where M is the length of the held-out validation data. In this way, we can emphasize one N-gram for certain segment of input speech if this segment have reliable clues and information about the domain of N-gram.

4. Experimental Results

4.1 Korean Broadcast News Recognizer

In this section, experimental results are reported on a Korean broadcast news speech recognition system. 300 hours of broadcast news reports from three different television stations in Korea (KBS, MBC, and SBS) were used for training the CDHMM acoustic model. A 13 dimension MFCC, delta, delta-delta feature vectors, global cepstrum mean subtraction, and vocal tract length normalization were applied. The lexicon consisted of 65,997 pseudo-morpheme entries [11]. In total, 223M pseudo-morphemes of two newspapers' (Chosun Ilbo, Donga Ilbo) texts, and 51M pseudo-morphemes from the broadcast news transcriptions were used to build the background Katz-smoothed 4-gram. Three episodes of news data were used as validation data to decide the threshold of the word level confidence measure and the divergence score threshold of the retrieved documents. The test speech data was selected to have a sufficient time gap from the training speech and text data. The two hours of test speech data was taken from two different broadcast news programs. The test speech was automatically partitioned into 39 individual stories using a heuristic rule and GMM. Each story is associate with a single topic and used as a unit of the language model update.

4.2 Adaptation Corpus Experiment

We compared the adaptation corpus collected by the language modeling approach with the adaptation corpuses which were collected by the simple tf-idf model, and the BM25 weighting scheme [12]. In each IR technique, the adaptation corpus was collected from the training data for the background 4-gram. The document that had a higher divergence score than the empirically decided threshold score was collected as the adaptation corpus. The criterion to determine the threshold of divergence score was the perplexity of the raw adapted model. The threshold was found where the perplexity of the raw adapted N-gram on validation data stopped decreasing. Katz-smoothed 4-gram was also used to train the raw adapted model just like the background 4-gram. To build the final adapted models for speech recognition experiments, the raw adapted 4-grams were interpolated with the background 4-gram using global static linear interpolation coefficients. The global static interpolation coefficient was estimated using validation data. For all adaptation corpuses, the interpolation coefficients for the background 4-gram were decided at approximately 0.8. Table 1 displays the results. According to Table 1, the bigram- and trigram-query based retrieval model show better word error rate (WER) performance than other IR techniques. The adapted 4-gram with trigram-query based IR yields a 7.5% improvement over the background 4-gram. The average number of documents of the retrieved adaptation corpus for one story is also represented in Table 1. One document consists of 11.8 sentences on average.

Table 1 Word error rates of speech recognition and perplexities (PPL) of final adaptation language models using various IR techniques. The size of adaptation corpus (AC) which is the average number of documents for one story is also represented.

		PPL	WER	AC size
Backgorund 4-gram		136.23	18.64	# of doc
IR Technique	Simple tf-idf	99.71	17.82	1382.3
	BM25 weighting scheme	89.04	17.56	1379.9
	Unigram-query based IR	92.40	17.83	1333.8
	Bigram-query based IR	88.94	17.49	1350.1
	Trigram-query based IR	84.18	17.25	1343.1

Table 2 Characteristics of the various sizes of the N-best lists and the performance of the corresponding dynamic interpolation coefficients.

	Background model rescored N-best list		Raw adapted model rescored N-best list		WER
N	Precision	Recall	Precision	Recall	
2	0.661	0.820	0.604	0.792	16.24
5	0.574	0.831	0.536	0.801	16.08
10	0.500	0.844	0.455	0.814	15.71
30	0.380	0.860	0.330	0.825	15.94
50	0.377	0.867	0.285	0.830	16.30
100	0.265	0.889	0.167	0.842	16.22

4.3 Dynamic Interpolation Coefficient Experiment

Experiments on the dynamic language model interpolation coefficients were carried out. The incoming test speech was partitioned by varying the size of the N-best list. Following this, the dynamic interpolation coefficients were estimated using the above mentioned segmented input speeches. The final adapted model was generated with these coefficients. Table 2 shows the recall and the precision of the various sizes of the N-best lists and the WER performance of final adapted 4-gram based on the corresponding coefficients. The recall and precision of each N-best list were measured. The recall is defined as the fraction of correct words which were recognized, and the precision is defined as the fraction of recognized words which were correct. A segmentation with 10-best list show of the best WER performance of the corresponding adapted 4-gram.

Next, we compared the proposed dynamic interpolation coefficient with the other interpolation coefficients. Three interpolation coefficients were implemented for a comparative experiment. First, the global static interpolation coefficient was estimated from the held-out validation data (held-out static) [2]. 0.8 was used as the static interpolation coefficient for the background language model. Second, The interpolation coefficient was defined as the function of history. The statistics about history were acquired from the held-out data (held-out history dependent) [5]. Finally, the on-line algorithm was used to obtain the dynamic interpo-

Table 3 Comparative experimental results with proposed dynamic interpolation coefficient and other interpolation coefficients.

The type of interpolation coefficient	WER
Held-out static	17.25
Held-out history dependent	17.40
On-line history dependent	16.59
Proposed segment dependent	15.71

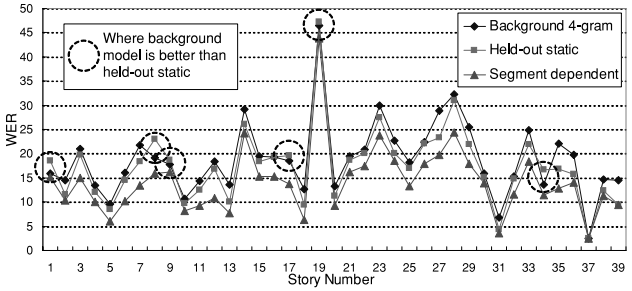

Fig. 2 Word error rate per story. the story is the unit of language model adaptation. Proposed method shows better performance than the background model consistently.

lation coefficient (on-line history dependent) [6]. Using this method, the previous hypothesis history determined the interpolation coefficient of the current word hypothesis while the decoding procedure. The results of the recognition experiments are shown in Table 3. The proposed segment dependent interpolation coefficient reduces the WER by 8.9% over the held-out static coefficient.

The performance of proposed adapted N-gram and background N-gram were then evaluated for each story of the test news data. As shown in Fig. 2, the proposed method consistently provided better performance than the background N-gram. However, the static interpolation coefficient showed a worse performance in six stories compared to the background model.

5. Conclusion

In this study, a method of N-gram adaptation using using the IR technique without additional text data was presented. The language modeling approach to IR with an N-gram retrieval model was found to reduce both the perplexity and the WER of the adapted model, which is based on other IR methods. Additionally, we proposed a dynamic interpolation coefficient that employs the word hypotheses of the input speech as held-out validation data. We showed that the final adapted N-gram with a dynamic interpolation coefficient always outperformed the case in which only background N-gram was used.

References

[1] J.R. Bellegarda, "An overview of statistical language model adaptation," Proc. ISCA Adaptation Methods in Automatic Speech Recognition, pp.165–174, Sophia Antipolis, France, Aug. 2001.
[2] B. Bigi, Y. Huang, and R. De Mori, "Vocabulary and language model adaptation using information retrieval," Proc. Int. Conf. on Spoken Language Processing, pp.1361–1364, Jeju, Korea, 2004.
[3] X. Zhu and R. Rosenfeld, "Improving trigram language modeling with the world wide web," Proc. Int. Conf. Acoustics, Speech, Signal Processing, pp.533–536, 2001.
[4] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to adhoc information retrieval," Proc. ACM SIGIR'01, pp.334–342, 2001.
[5] R. Kneser and V. Steinbiss, "On the dynamic adaptation of stochastic language models," Proc. Int. Conf. Acoustics, Speech, Signal Processing, vol.2, pp.586–589, 1993.
[6] A. Kalai, S. Chen, A. Blum, and R. Rosenfeld, "On-line algorithms for combining language models," Proc. Int. Conf. Acoustics, Speech, Signal Processing, vol.2, pp.745–748, Phoenix, USA, 1999.
[7] L. Mangu, Finding consensus in speech recognition, Ph. D. Thesis, Johns Hopkins Univ., 2000.
[8] D.H. Miller, T. Leek, and R. Schwartz, "A hidden Markov model information retrieval system," Proc. ACM SIGIR 1999, pp.214–221, 1999.
[9] P. Ogilvie and J. Callan, "Experiments using the Lemur toolkit," Proc. 10th Text REtrieval Conference, pp.103–108, 2001.
[10] F. Jelinek, "Self-organized language modeling for speech recognition," Readings in Speech Recognition, pp.450–506, Morgan Kaufmann, 1989.
[11] O.W. Kwon and J. Park, "Korean large vocabulary continuous speech recognition with morpheme-based recognition units," Speech Commun., vol.39, pp.287–300, Feb. 2003.
[12] S.E. Robertson, S. Walker, M.H. Beaulieu, M. Gatford, and A. Payne, "Okapi at TREC-4," Proc. Fourth Text REtrieval Conf. (TREC-4), pp.73–96, Oct. 1996.