A Rapid Bayesian Adaptation of N-gram Language Models Using Cross-word Correlation

Hui Jiang, Keikichi Hirose, Nobuaki Minematsu, Koki Sasaki and Takaaki Moriya

Technical Report CS-2005-02

Feb. 4, 2005

Department of Computer Science and Engineering

4700 Keele Street North York, Ontario M3J 1P3 Canada

A Rapid Bayesian Adaptation of N-gram Language Models Using Cross-word Correlation

Hui Jiang[*], *Keikichi Hirose*[‡], *Nobuaki Minematsu*[†], *Koki Sasaki*[†] *and Takaaki Moriya*[‡]

[*] Department of Computer Science, York University,

4700 Keele Street, Toronto, Ontario M3J 1P3, CANADA

[†] Department of Information and Communication Engineering,

[‡] Department of Frontier Informatics, School of Frontier Sciences

The University of Tokyo, Hongo 7-3-1, Bunkyu-ku, Tokyo 113-0033, JAPAN

Correspondence:

Prof. Hui Jiang, Department of Computer Science, York University,

4700 Keele Street, Toronto, Ontario M3J 1P3, CANADA

Phone: (416)736-2100 x33346 (Hui Jiang) **Fax:** (416)736-5872 (Hui Jiang)

Email: hj@cs.yorku.ca {koki,hirose,mine,moriya}@gavo.t.u-tokyo.ac.jp

Abstract

In this paper, we study a fast adaptation problem of n-gram language models under the MAP estimation framework. We propose a heuristic method to explore cross-word correlation to accelerate the MAP adaptation of n-gram models. According to the correlation, occurrence of one word in adaptation text can be used to predict all possible n-grams which will likely appear in the same adaptation text. Then the predicted occurrence is incorporated into the MAP estimation of n-gram models. In this way, a large n-gram model can be efficiently adapted with only a small amount of adaptation data. We have conducted two experiments to evaluate the proposed fast adaptation technique, e.g., topic adaptation within a domain and cross-domain adaptation. All experimental results clearly show that the proposed fast adaptation approach is very efficient and effective to adapt a large n-gram model to a new task quickly, in terms of perplexity reduction and speech recognition improvements. It is also shown that the proposed fast adaptation technique significantly outperforms the conventional MAP adaptation, especially when we have very limited amount of adaptation data.

Index Terms:

Language Model Adaptation, Bayesian Learning, Maximum *a posteriori* (MAP) Estimation, N-gram, Cross-word Correlation, Automatic Speech Recognition (ASR)

1 Introduction

It is well known that a proper statistical language model (LM) plays an important role in large vocabulary automatic speech recognition (LVASR) system. A statistical language model is used to calculate probability of observing any given word sequence (sentence). The most popular LM in speech recognition is n-gram modeling, which assumes the probability of observing a word only depends on its preceding $n-1$ words ($n-1$-order Markovian assumption) in a sentence. Based on information provided by an n-gram language model, an LVASR system is able to reduce search space considerably and successfully resolve the serious confusion of speech signals in acoustic level. But, estimating a reliable n-gram model usually requires a huge amount of text corpus. And it is always very expensive and tedious to collect a lot of text data for many different domains. Thus, it is not practical to collect enough text data to train a domain-dependent n-gram model for every specific task. Today's LVASR systems always use a general-purpose task-independent (TI) n-gram language model for all tasks or domains. However, because language characteristics usually differ more or less from one domain to another and even from topic to topic within a domain, it is strongly desirable to have a task-dependent (TD) n-gram model for each specific task in order to achieve an optimal recognition performance for that particular task. Obviously, one feasible strategy to achieve this goal is to adopt adaptive learning method, i.e., we adapt a TI n-gram model to each target task by using merely a small amount of text data collected for that task. In the literature, a bunch of research works have been done to adapt n-gram language models in the area of speech recognition. Generally speaking, three different methods have been proposed to perform n-gram language model adaptation.

The first category is called *interpolation*-based methods. In this category, usually several n-gram models are first built to represent different knowledges or domains. Then a linear interpolation of all available n-gram models is calculated as final language model for speech recognition, such as in [Kuhn & De Mori, 1990], [Jelinek et al, 1991], [Matsunaga et. al, 1992], [Kneser & Steinbiss, 1993], [Clarkson & Robinson, 1997], [Iyer & Ostendorf, 1999], [Kalai et al, 1999], and etc.. In [Kuhn & De Mori, 1990], and [Clarkson & Robinson, 1997], a so-called *cache* model, which is calculated from the most recently observed text data on the fly, is interpolated with a general n-gram model to

track the article level correlation. In [Jelinek et al, 1991] and [Matsunaga et. al, 1992], a general topic-independent LM is adapted by linearly combining with a simple target topic related model. In [Kneser & Steinbiss, 1993] and [Iyer & Ostendorf, 1999], interpolation weights are dynamically updated from the observed history text based on maximum likelihood (ML) criterion to enhance the interpolation-based LM adaptation. The interpolation method is very simple and has been reported to be quite effective by many different sites.

Secondly, another major work related to LM adaptation is based on *Minimum Discrimination Information* (MDI) [Della Pietra et al, 1992, Rao, Monkowski & Roukos, 1995, Rao, Dharanipragada & Roukos, 1997]. In MDI-based adaptation method, the new LM is calculated by minimizing discrimination information (Kullback-Liebler distance) from an original TI model under some constraints which reflect important features of text data in target topic. Usually an iterative algorithm, called *generalized iterative scaling* (GIS)[Darroch & Ratcliff, 1972], is used to solve the constrained optimization problem for n-gram models. A drawback of MDI-based adaptation is its computation complexity because the iterative scaling has to be done for all elements in an n-gram model. If an n-gram model's size is large, the adaptation procedure becomes quite slow.

Thirdly, mainly motivated by the success of Bayesian learning for HMM[Gauvain & Lee, 1994], some researchers have proposed to adapt n-gram language models under the Bayesian framework, especially the *maximum a posteriori* (MAP) estimation[Federico 1996] [Masataki et al, 1997]. Although the MAP estimation of n-gram models results in a very simple formula, a major problem with the MAP estimation of n-gram models is that it is very slow to converge in terms of learning rate, which means that lots of adaptation data are needed to make adaptation effective. In fact, not only particular to the Bayesian method, this problem is also general for all other above-mentioned n-gram LM adaptation methods. Therefore, it becomes a very interesting question how to perform rapid adaptation for n-grams, namely efficiently update n-gram models from only a very small amount of adaptation data. Although we have many fast adaptation techniques available for acoustic models, e.g., HMM's, such as transformation-based methods[Leggetter & Woodland, 1995], we have not yet found any good solution for n-gram LM adaptation because n-gram models are even harder to handle in some senses, e.g., an n-gram model is flat and contains a large number of parameters, and also these pa-

rameters are mutually constrained, which makes it difficult to apply transformation-based adaptation strategy to n-gram models.

In this paper, we study the problem how to rapidly adapt n-gram models from a Bayesian viewpoint. As we have mentioned, the MAP (maximum a posteriori) estimation of n-gram models has a straightforward form to implement, but it is too slow to converge in a new task domain. In this work, we focus on a rapid adaptation method for n-gram models under the Bayesian learning framework. Starting from the MAP formulation of n-gram models, we propose a heuristic method to explore correlation between each key-word and all n-grams appearing in a near context. And the recorded correlation information is utilized to make the MAP adaptation of n-gram models faster, more efficient and effective. Particularly, when we estimate task-independent n-gram model from a large text corpus, we also explore and record all information about the correlation between every key-word and its surrounding n-grams in a correlation matrix. Each element of the matrix represents frequency of their co-occurring in a single document. When we adapt the task-independent n-gram model to a certain target task with only a small amount of adaptation text, appearance of any key-word in adaptation text can be used to predict occurrence of all other possible n-grams based on the correlation matrix. Then the predicted occurrence of each n-gram is integrated with its actual occurrence. Finally, a task-adaptive n-gram model is derived under the framework of MAP estimation. In this way, a large n-gram language model can be rapidly updated based on merely a very small amount of task-dependent text data. In this work, the fast adaptation technique is evaluated in two LM adaptation tasks, namely topic adaptation within a domain and cross-domain adaptation. All experimental results clearly show that the approach works pretty well to adapt n-gram language models into a new task in terms of both perplexity reduction and speech recognition performance improvement. When comparing with other conventional methods, the results also show that it significantly outperforms the standard MAP adaptation, especially in case only a very limited amount of adaptation data is available. As an example, in cross-domain adaptation case, the proposed approach can reduce perplexity of a relatively large bi-gram model nearly 30% (from 210 down to 152) with only 133 adaptation sentences from the target domain while a widely-used MAP adaptation can improve about 20% in the same case.

In this paper, we summarize and gather together all results scattered in [Sasaki, 2000, Sasaki, Jiang & Hirose, 2000, Moriya et al, 2001, Hirose, Minematsu, Moriya, 2002] to make our works more accessible to general readership. The remainder of this paper is organized as follows. In section 2, we first introduce the standard MAP adaptation formulation for n-gram language models. Next, in section 3, we present our new fast adaptation method and give all details on how to explore cross-word correlation in text to enhance the MAP adaptation for n-gram models. Then we report within-domain topic adaptation experiments in section 4 and cross-domain LM adaptation experiments in section 5 respectively. Finally, we conclude the paper with our findings and some possible future works in section 6.

2 MAP Estimation of N-gram Model

Today n-gram models have become the dominant statistical language modeling method for large vocabulary speech recognition. In n-gram modeling, any given word sequence (sentence) $S = \{w_1 w_2 \cdots w_T\}$ is assumed to be a Markov chain, observing the current word w_i only depends on its immediate history h_i. Thus, probability of observing S is calculated as

$$P(S) = P(w_1 w_2 \cdots w_T) = \prod_{i=1}^{T} P(w_i \mid h_i) \tag{1}$$

Depending on the case, history h_i could be the preceding word w_{i-1} (in bigram), the preceding two words $w_{i-2}w_{i-1}$ (in trigram), or even a longer segment.

Generally speaking, an n-gram language model Λ is composed of a set of word occurrence conditional probabilities on the corresponding histories, i.e., $P(w|h)$. If we denote $P(w|h) \equiv \lambda_{hw}$, the n-gram model Λ can be expressed as

$$\Lambda = \{\ \lambda_{hw} \mid w \in W \text{ and } h \in H\ \}. \tag{2}$$

where W denotes the set of all words in vocabulary and H all possible histories. Obviously, the n-gram model parameters λ_{hw} follow the constraint

$$\sum_{w \in W} \lambda_{hw} = 1 \tag{3}$$

for every h in H.

Given any text corpus $\mathbf{T} = w_1 w_2 ... w_n$, the likelihood function of n-gram model Λ is computed from eq.(1) as

$$\begin{aligned} l(T|\Lambda) &= P(w_1 w_2 ... w_n) = \prod_{i=1}^{n} P(w_i|h_i) \\ &= \prod_{w \in W} \prod_{h \in H} \lambda_{hw}^{N_{hw}} \end{aligned} \tag{4}$$

where N_{hw} denotes occurrence number (frequency) of n-gram hw in text T. By taking the constraints in eq.(3) into account, the maximum likelihood (ML) estimation of n-gram model which maximizes $l(T|\Lambda)$ can be derived as:

$$\lambda_{hw}^{(ML)} = \frac{N_{hw}}{\sum_{w \in W} N_{hw}}. \tag{5}$$

From eq.(4), we note that the likelihood function of n-gram model is a multinomial distribution. Its natural conjugate prior is the so-called *Dirichlet* distribution[DeGroot, 1970]:

$$p(\Lambda) = p(\{\lambda_{hw}\}) \propto \prod_{w \in W} \prod_{h \in H} \lambda_{hw}^{\alpha_{hw}-1} \tag{6}$$

where $\{\alpha_{hw} \mid h \in H, w \in W\}$ are hyperparameters, which usually are estimated from a task-independent corpus T^i:

$$\alpha_{hw} = N_{hw}^i + 1 \;\; (w \in W, \; h \in H) \tag{7}$$

where N_{hw}^i denotes occurrence number of hw in the corpus T^i.

According to Bayes' theorem, given an adaptation text data T^a, the posterior pdf is

$$p(\Lambda|T^a) \propto p(\Lambda) \cdot l(T^a|\Lambda) \propto \prod_{w \in W} \prod_{h \in H} \lambda_{hw}^{N_{hw}^i + N_{hw}^a} \tag{8}$$

Then the MAP (*maximum a posteriori)* estimation of an n-gram model is obtained as:

$$\Lambda^{(MAP)} = \arg\max_{\Lambda} \; p(\Lambda|T^a) = \arg\max_{\Lambda} \; p(\Lambda) \cdot l(\Lambda|T^a) = \arg\max_{\Lambda} \prod_{w \in W} \prod_{h \in H} \lambda_{hw}^{N_{hw}^i + N_{hw}^a} \tag{9}$$

Similarly, by considering the constraints in eq.(3), the MAP estimation of n-gram model, to say λ_{hw}, can be derived as:

$$\lambda_{hw}^{(MAP)} = \frac{(N_{hw}^i + N_{hw}^a)}{\sum_{w \in W} (N_{hw}^i + N_{hw}^a)} \tag{10}$$

In the case of language model adaptation, the amount of adaptation data usually is much less than that of TI training data used to estimate the priors (or baseline language

model), i.e., $N^i_{hw} >> N^a_{hw}$. In many cases, in order to improve performance of the MAP adaptation, a widely-used method is shown as follows:

$$\lambda^{(MAP)}_{hw} = \frac{(N^i_{hw} + \beta \cdot N^a_{hw})}{\sum_{w \in W} (N^i_{hw} + \beta \cdot N^a_{hw})} \tag{11}$$

where a boosting factor β ($\beta \geq 1$) is used to simply boost N^a_{hw} to make it comparable with N^i_{hw}, and its proper value depends on relative amount difference of training and adaptation data on a case-by-case basis.

From the above eqs. (10) and (11), we note the MAP estimation of n-gram model has a very straightforward formulation. However, it converges too slow, even after the simple boost scaling scheme. Comparing with TI text data T^i, the adaptation text T^a usually has much less amount of data. Therefore, the values of N^a_{hw} will be much smaller than that of N^i_{hw}. A small amount of adaptation data will not change the value of λ_{hw} too much. So it usually requires a relatively large amount of adaptation data to make MAP-based adaptation effective. Because it is relatively costly to collect data in practice, it is strongly desirable to have a fast method to adapt n-gram model more efficiently.

3 Rapid MAP Adaptation Using Cross-word Correlation

Motivated by the MAP estimation for jointly correlated Gaussian mean vectors [Lasry & Stern, 1984] and correlated continuous density HMM[Huo & Lee, 1998], we are interested in studying how to accelerate Bayesian learning of n-gram models by using correlation information. However, it seems extremely difficult to directly model correlation between n-gram model's parameters in a rigid way. Alternatively, in this paper, we consider to utilize cross-word correlation empirically. In other words, if two words or n-grams tend to frequently appear together in a close context, we will think these two words are strongly correlated. When one of them occur in a document, it indicates another one and causes its probability estimation to change. This kind of correlation information has been investigated in the so-called *trigger* language model in [Lau, Rosenfeld & Roukos, 1993, Rosenfeld, 1996] to capture information from a long distance history in statistical language modeling, where the correlation is cast as a set of

feature constraints. These constraints are expressed in terms of marginal distribution. Then the information is incorporated into n-gram language models based on the maximum entropy principle. In contrast to [Lau, Rosenfeld & Roukos, 1993, Rosenfeld, 1996], in this work, we attempt to incorporate the correlation information into n-gram language modeling in context of model adaptation under a Bayesian framework, which results in a much simpler solution than the Maximum Entropy estimation.

3.1 Basic Adaptation Algorithm

In this section, starting from the MAP formulation of n-gram models, we propose a heuristic method to investigate cross-word correlation to make the Bayesian adaptation of n-gram model fast, efficient and effective. Concretely, when we estimate a task-independent n-gram model from a large text corpus, we also explore and record all information about correlation between any a key-words and all its surrounding n-grams in the corpus, i.e. a correlation matrix to indicate the frequency of co-occurrence in a close context. When it is needed to update the task-independent n-gram model to a certain target task based on a small amount of adaptation text, by using the estimated correlation matrix, appearance of one certain key-word in adaptation text can be used to predict occurrence of all other n-grams. Then all of these predicted occurrences are added with the actual occurrence in the adaptation text. Finally, a task-adaptive n-gram model is derived under the framework of MAP estimation. In this way, a large n-gram language model can be rapidly updated based on merely a small amount of task-dependent text data.

Our proposed fast adaptation algorithm is performed as follows:

A. Estimate TI n-gram model and record correlation matrix:

1. From the task-independent corpus T^i, we estimate a TI n-gram model $\{\lambda_{hw}\}$ and record all sufficient statistics N^i_{hw} for every n-gram hw.

2. Based on the TI data T^i, we build a common word list (CWL) which includes all common words appearing equally everywhere in a language, such as prepositions, adverbs, auxiliary verbs, and etc. Hereafter, a *key-word* is defined as any word not included in the CWL.

3. Partition the TI data T^i into some contiguous segments: $T^i = T_1^i T_2^i ... T_K^i$. Hereafter, we call each segment $T_k^i (1 \leq k \leq K)$ as a *document*. Here each document T_k^i can be a sentence, a paragraph, or even an article. Although the partition is flexible, it is better to make sure all sentences in a document are from a single domain (or topic).

4. For any n-gram hw and every document T_k^i $(1 \leq k \leq K)$, we calculate its co-occurrence indicator with all key-word v appearing in T^i, i.e.,

$$q_{v[hw]}^k = \begin{cases} 1 & \text{if } v \text{ and } hw \text{ co-occur in } T_k^i, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

   Then we summarize $q_{v[hw]}^k$ over all documents T_k^i to obtain a correlation matrix between any a key-word v and any n-gram hw, i.e.,

$$q_{v[hw]} = \sum_{1 \leq k \leq K} q_{v[hw]}^k \tag{13}$$

   Note that in this paper we don't consider the unknown word problem in adaptation for simplicity. Thus we just set $q_{v[hw]} = 0$ when either hw or v contains an unknown word. All unknown words are denoted as *UNK*.

B. Fast adaptation

5. Given adaptation data T^a, we first collect sufficient statistics, N_{hw}^a, i.e., occurrence number in T^a, for every n-gram hw.

6. According to the correlation matrix $q_{v[hw]}$, we compute a predicted occurrence number Q_{hw} for every n-gram hw based on all key-words v occurring in T^a. That is,

$$Q_{hw} = \sum_{v \in T^a, v \overline{\in} CWL} N_v^a \cdot q_{v[hw]} \quad (\text{for all } hw) \tag{14}$$

   where N_v^a denotes occurrence number of word v in T^a.

7. The predicted occurrence of an n-gram hw is added to its actual occurrence. Based on the MAP formulation, we estimate a new n-gram model as follows:

   - When $w \neq UNK$ (w is a known word),

$$\lambda'_{hw} = \frac{N_{hw}^i + \beta \cdot N_{hw}^a + \alpha \cdot Q_{hw}}{\sum_{w \in W} N_{hw}^i + \beta \cdot \sum_{w \in W} N_{hw}^a + \alpha \cdot \sum_{w \in W} Q_{hw}} \tag{15}$$

where α is a weight to control the contribution of the predicted occurrence number.

- When $w = UNK$,[1] (w is an unknown word)

$$\lambda'_{hw} = \frac{N^i_{hw} + \beta \cdot N^a_{hw}}{\sum_{w \in W} N^i_{hw} + \beta \cdot \sum_{w \in W} N^a_{hw}} \tag{16}$$

8. Obviously, the new n-gram model updated from eqs.(15) (16) does not satisfy the constraint $\sum_{w \in W} \lambda'_{hw} = 1$. Thus we use the following strategy to normalize λ'_{hw} and finally get the adapted n-gram model as:

$$\lambda_{hw} = \begin{cases} C(h) \cdot \lambda'_{hw} & w \neq UNK, \\ \lambda'_{hw} & w = UNK \end{cases} \tag{17}$$

where

$$C(h) = \frac{\sum_{w \in W, w \neq UNK} \lambda''_{hw}}{\sum_{w \in W, w \neq UNK} \lambda'_{hw}} \tag{18}$$

and λ''_{hw} denotes the pure MAP estimate:

$$\lambda''_{hw} = \frac{N^i_{hw} + \beta \cdot N^a_{hw}}{\sum_{w \in W} N^i_{hw} + \beta \cdot \sum_{w \in W} N^a_{hw}} \tag{19}$$

3.2 Some Implementation Issues

In step 2 of the algorithm, we need filter out those common words which equally appear everywhere in various domains. In this work, we do not select particular trigger pairs based on mutual information as in [Rosenfeld, 1996]. Instead, for simplicity, we assume every key-word uniformly "triggers" all n-grams in the same document. In this case, it is crucial to remove these common words from correlation consideration. Otherwise, predicted occurrence of n-grams will be over-estimated.

In step 4, we calculate a correlation matrix $q_{v[hw]}$ between all key-words and all appearing n-grams, which are saved for fast adaptation. A large value of $q_{v[hw]}$ indicates that v and hw co-occur in a single document quite frequently, which implies they are highly correlated. In eq.(13), an un-normalized correlation is computed. Alternatively, as

[1] In case w is unknown word, in eq.(15), we have $Q_{hw} = 0$ in numerator but a large number $\sum_{w \in W} Q_{hw} \geq 0$ in denominator. Thus eq.(15) will underestimate the probabilities of UNK.

in [Moriya et al, 2001], we can also calculate a normalized correlation matrix as:

$$\bar{q}_{v[hw]} = \frac{1}{K} \sum_{1 \le k \le K} q^k_{v[hw]}. \tag{20}$$

In this case, every matrix element $\bar{q}_{v[hw]}$ becomes a number between 0 and 1.

In the above algorithm, we only consider the correlation in task-independent training set T^i. Following steps 3 and 4, we can calculate an on-line version of correlation matrix $q^a_{v[hw]}$ from adaptation data T^a as well. Then we can similarly predict occurrence of n-gram hw according to $q^a_{v[hw]}$ as:

$$Q^a_{hw} = \sum_{v \in T^a, v \overline{\in} CWL} N^a_v \cdot q^a_{v[hw]} \ \ \text{(for all } hw\text{)}. \tag{21}$$

The predicted number Q^a_{hw} can be similarly included into the MAP estimation formula. If w is not an unknown word, then we have

$$\lambda'_{hw} = \frac{N^i_{hw} + \beta \cdot N^a_{hw} + \alpha \cdot Q_{hw} + \lambda \cdot Q^a_{hw}}{\sum_{w \in W} N^i_{hw} + \beta \cdot \sum_{w \in W} N^a_{hw} + \alpha \cdot \sum_{w \in W} Q_{hw} + \lambda \cdot \sum_{w \in W} Q^a_{hw}} \tag{22}$$

where λ is another control parameter. Finally, we follow step 8 to normalize λ'_{hw} to obtain the adapted n-gram model.

4 Experiments(I): Topic Adaptation

In the first set of experiments, the proposed fast adaptation method is evaluated in the task of topic adaptation within domain. We choose Japanese Mainichi newspaper as the domain and all articles related to the "gulf war" are selected as the target topic. In the experiment, totally 5000 articles (approximately 1.2M words) from 1991 electronic Mainichi newspaper are used as topic-independent(TI) corpus (excluding all articles related to the "gulf war"). Among all articles related to the "gulf war", we choose 100 sentences as our evaluation set to estimate perplexity of different language models, and several adaptation sets with different sizes, including 100, 300, 500, 750, 1000 sentences respectively. More details about the text corpus can be found in Table 1. The baseline bi-gram model is estimated from TI training corpus with the standard method[Clarkson & Rosenfeld, 1997]. Good-Turing discounting method is used in bigram model construction. And vocabulary size of bigram model is chosen as 5k. Then this initial baseline bi-gram model is adapted

to the target topic with different amount of adaptation data by using the rapid Bayesian learning method proposed in this paper. Then the adapted language models are compared with the initial baseline model and other adapted models from the standard MAP adaptation, in terms of perplexity improvement and speech recognition error reduction. In all experiments of this paper, we use CMU-Cambridge statistical language modeling toolkit to build baseline bi-gram models and calculate perplexity.

4.1 How to build Common Word List (CWL)

As we have mentioned, it is important to filter out some common words to avoid overestimating cross-word correlation. Following the work in [Kawahara & Doshita, 1999], we use mutual information to select the most common words for CWL. We use all articles in 1991 Mainichi newspaper (except those on 'Gulf War') and partition them into 10 topics, i.e., $T = \{t_1, t_2, ..., t_{10}\}$. The mutual information of each word w and the text T is calculated as

$$I(T; w) = -\sum_{i=1}^{10} P(t_i) \log P(t_i) + \sum_{i=1}^{10} P(t_i|w) \log P(t_i|w) \tag{23}$$

where $P(t_i) = 0.1$ and $P(t_i|w) = \frac{\text{frequency of } w \text{ in } t_i}{\text{frequency of } w \text{ in } T}$. Thus, $I(T; w)$ indicates nonuniformity of frequency of the word w in various topics.

At first, we select top 20000 words according to their frequencies in T. Then we calculate $I(T; w)$ for each word in the top list and sort them according to their $I(T; w)$ values. Finally we pick up the last N_c words in the sorted list which have smaller $I(T; w)$ values to build a Common Word List (CWL) of size N_c.

4.2 Effects of partition unit and α (as in eq. (22))

In step 3 of the algorithm, we need to partition TI training corpus into many small segments to calculate correlation between words. Here we investigate what partition unit is proper for our fast adaptation method. We partition the TI training set based on three different partition units, namely sentence (PS), paragraph (PP) or article (PA). In figure 1, perplexity calculated in the evaluation set is shown as a function of α (as in eq. (15) or (22)) for different partition units, where we fix $\beta = 1$ and $\lambda = 0$. From the experimental results, we see that for all three different partition units the rapid adaptation method

can achieve significant perplexity improvements over the baseline model (as well as the model trained on adaptation data only) once we select a proper value for α. We conclude that PS and PP give much more perplexity reduction than PA, and PS yields the best performance. In the following experiments, we will use PS as the default partition unit. As for α, it is definitely task-dependent and but usually gives good performance in $[0.01, 0.1]$. In our following experiments, we fix $\alpha = 0.03$ except explicitly stated.

4.3 Effect of the size of CWL (N_c)

In this section, we will investigate the influence of CWL size, N_c, in perplexity reduction in our fast adaptation approach. We fix $\alpha = 0.03$ and partition TI training corpus by sentence. In Figure 2, we show perplexity as a function of N_c in two cases, namely using adaptation set C (500 sentences) and adaptation set D (750 sentences) for adaptation. From the results in Figure 2, we can see the fast adaptation method performs reasonably well in a quite wide range of N_c, i.e.,$[6000, 16000]$. In particular, we choose $N_c = 7000$ in the following topic adaptation experiments.

4.4 Adaptation Performance Measured in Perplexity

In the experimental results reported in previous sections, we only consider the correlation calculated from TI training data T^i, e.g., just using eq.(15). We denote it as FA method in the following. To further accelerate adaptation, we can compute correlation Q^a_{hw} from adaptation data T^a. In other words, we replace eq.(15) with eq.(22) in our rapid adaptation algorithm, which is denoted as FA2 method. Obviously, FA is a special case of FA2 with $\lambda = 0$. In Figure 3, we compare our fast adaptation methods FA, FA2 with the standard MAP adaptation (eq. 10) for various amount of adaptation text data. It is clear that by using correlation between key-words and the corresponding n-grams both FA and FA2 converge much faster than MAP, and FA2 with $\beta = 30$ gives the maximum perplexity reduction. It is amazing that our rapid adaptation method FA2 can reduce perplexity of bi-gram model in target topic about 20% (from 102 down to 83) with only 100 sentences from the target topic, and nearly 35% (from 102 to 67) with 500 sentences. From the results, we also see that the methods without the boost scaling ($\beta = 1$) gives a very poor performance. Thus, in all following experiments, we always use a proper boost

scaling factor β in both the baseline MAP estimation and the FA2 method.

4.5 Speech Recognition Results (I)

As the most important application of n-gram models, in this section, we will conduct some speech recognition experiments to see how much the fast LM adaptation method will help speech recognition. We first collect some speech data on the target task domain from one male speaker. The speaker is asked to read all sentences in the evaluation set. In decoding, we employ a standard Japanese JULIUS decoding software and a Japanese state-tied triphone acoustic model (totally 3k distinct states and 16 Gaussian mixtures per state) supplied with the decoder[Kawahara et al, 2000]. During recognition we use the same recognizer setup except different bi-gram language models derived from various adaptation methods. Comparative recognition results in terms of word accuracy[2] are shown in Figure 4. We have found that our fast language adaptation method can generally improve speech recognition performance even when we have only 100 or 300 sentences to adapt language model and it significantly outperform the standard MAP adaptation. By using 100 sentences to adapt language model, we can improve recognition word accuracy 3.7% in absolute. Because the recording condition of test speech data seriously mismatches with the available acoustic models, unfortunately the absolute performance of the baseline speech recognition system is too low, around 40%. In the next section, we will report more speech recognition results on a better baseline system.

5 Experiments(II): Cross-Domain Adaptation

In the second set of our experiments, we attempt to investigate the effect of our fast adaptation algorithm in case of cross-domain adaptation, where we adapt a task-independent (TI) language model to a significantly different domain. Here, we still train a TI bi-gram model from the same Mainichi Japanese newspaper corpus, but using a much larger amount of text data. As the target domain, we choose Japanese translation of a famous novel, name *Peter Pan* originally written by J.M. Barrie. The text data used in this part of experiments is summarized in Table 2. Based on TI training corpus, we use

[2]The word accuracy is defined as 1-WER, WER denotes word error rate.

CMU-Cambridge statistical language modeling toolkit to build a standard bi-gram model consisting of totally 20k words. In our experiments, the standard MAP LM adaptation and the proposed fast LM adaptation using cross-word correlation are used to adapt the baseline bi-gram model to the target domain based on three sets of adaptation data, which all comes from the novel *Peter Pan* and contains 133, 526 and 699 sentences respectively. At last, language models are evaluated in a disjointed evaluation set from the target domain. We know, in cross-domain adaptation, two issues need to be addressed, namely re-estimating word probabilistic distribution and dealing with out-of-vocabulary (OOV) words. In this work, we only concentrate on the first issue, e.g., word probability re-estimation. The OOV rate in the evaluation set remains quite huge, about 8%. As for the other experimental setups, in this part we try to maintain the best configuration we have achieved in section 4. For example, we partition training corpus in the unit of sentence to estimate correlation matrix. And, we use the same mutual information based method to select common word list (CWL) and choose its size $N_c = 8000$ in this part of experiments.

5.1 Evaluation in Perplexity

In the experiments, we first examine the standard MAP adaptation with the boosting factor β, e.g. eq.(11). We have found that the optimal value for the boosting factor β is around $[1000, 3000]$. In the following, we simply fix $\beta = 2500$. We note the optimal value for β in this part of experiments is much larger than what we used in section 4. The reason is that we have a much larger training corpus here and thus need a larger β to boost N^a_{hw} to make it comparable with N^i_{hw}. Another major difference in this part is that we use the normalized method, e.g., eq.(20) instead of eq.(13), to calculate the correlation matrix $q_{v[hw]}$. Two choices give very similar performance but cause a huge difference in value range of α and λ. In this case, we find that a suitable value for α is around $1e10$ and $1e7$ for λ.

In Figure 5, we compare our best fast adaptation method FA2 with the standard MAP adaptation (with boosting) in case of various amount of adaptation data. From the results, we can see, when we test the baseline bigram model (trained from newspaper corpus) on the evaluation set (*Peter Pan*), perplexity is very high, about 210, which

clearly shows the difference between two domains. When we have only 133, 526 and 699 adaptation sentences available, the standard MAP (with boosting) reduces the perplexity down to 170, 128 and 105 while FA2 can improve further to 152, 113 and 96, respectively. Even in cross-domain adaptation, the FA2 method still significantly outperforms the MAP adaptation in terms of perplexity reduction when we only have very limited amount of adaptation data.

5.2 Speech Recognition Results (II)

Similar as in section 4.5, we ask one male speaker to read the evaluation set and record a small speech database for speech recognition experiment. Here we still use the same decoder, JULIUS, and its accompanied acoustic models. We compare speech recognition performance based on the original LM, MAP-adapted LM's and FA2-adapted LM's. Experimental results are shown in Figure 6. When we conduct speech recognition experiments based on the original baseline bigram model, it gives word accuracy 64.8%. When we adapt LM to the target domain, recognition performance can be improved significantly. For MAP adaptation (with boosting), word accuracy goes up to 66.7%, 68.4% and 70.9% when we use 133, 526, 699 adaptation sentences, respectively. And in all three situations, the proposed FA2 method yields a consistent improvement over MAP, i.e., 67.6%, 69.9% and 71.8%, respectively.

6 Conclusions

Since a speech recognition system can be deployed for various task domains and language characteristics significantly differ from one case to another, it is strongly desirable to build statistical language models, mainly n-gram models, in an adaptive mode just as what we have done for acoustic models, namely HMM's. We know the MAP estimation[Gauvain & Lee, 1994] has been shown as a good tool to adapt statistical models. However, when using the MAP method to adapt n-gram models, we encounter the serious problem that it converges too slow in a new task domain. In this paper, we have proposed a heuristic method to accelerate the MAP adaptation of n-gram language models based on cross-word correlation. To evaluate the proposed method, we have carried

out two sets of LM adaptation experiments: i)topic adaptation within domain; ii) cross-domain adaptation. In all experiments, the new method is compared with a widely-used MAP adaptation method in terms of both perplexity reduction and speech recognition improvement. All experimental results clearly show that our novel method significantly outperforms the traditional MAP adaptation, and a significant perplexity reduction of n-gram models in target domain has been observed even when we only have few hundred adaptation sentences from the domain. And some ASR experimental results also show we can consistently improve word accuracy about a few percentage by adapting a task-independent language model with only a few hundred sentences of adaptation text from the target domain. It is known that fast adaptation techniques for n-gram language models is very important and useful in many practical applications. The work in this paper shows that using cross-word correlation in a language is a promising way to perform rapid n-gram language model adaptation. Also as shown in the paper, the MAP estimation is an attractive tool to incorporate cross-word correlation into n-gram language model estimation for adaptation purpose. The resultant solution is much simpler in terms of computational complexity and implementation convenience than other methods which use similar correlation in language modeling, such as trigger model based on maximum entropy[Lau, Rosenfeld & Roukos, 1993, Rosenfeld, 1996].

As far as future works are concerned, there are many modifications and enhancements we can do over the current version of this fast adaptation technique. For example, in this work, when we consider correlation between words and n-grams, we adopt a very simple method, i.e., we exclude a list of common words and consider every word not included in this list will uniformly "triggers" all n-grams in its near context. Obviously, the concept of a "trigger" pair in [Rosenfeld, 1996] can be used to define what n-grams will be "triggered" by each key-word in different context to model cross-word correlation in a better way. Besides, in this paper, we only consider the case where a word trigger an n-gram. Of course, all other combinations can be similarly tried, such as a word triggers a word, an n-gram triggers an n-gram, an n-gram triggers a word, etc. Moreover, in this work, we calculate the correlation matrix $q_{v[hw]}$ by using a binary value (0 or 1). Apparently, a probabilistic function depending on context can be used instead to capture more context information. At last, throughout all experiments in this paper, bi-gram

models are used to demonstrate efficacy of the new method for simplicity. Although this approach is equally applicable to all n-gram models in theory, it is very interesting to examine how it works for other n-gram models ($n \geq 3$).

References

[Bahl et al, 1989] Bahl, L.R., Brown, P. F., De Souza, P.V. & Mercer, R. L. (1989). A Tree-based statistical language model for natural language speech recognition. IEEE Trans. on Acoustic, Speech, and Signal Processing 37(7), 1001-1008.

[Clarkson & Rosenfeld, 1997] Clarkson, P. & Rosenfeld, R. (1997). Statistical language modeling using CMU-Cambridge toolkit. Proc. of European Conference on Speech Communication and Technology (Eurospeech'97), 2707-2710.

[Clarkson & Robinson, 1997] Clarkson, P. & Robinson, A. (1997). Language model adaptation using mixtures and an exponential decaying cache. Proc. of ICASSP'97, 799-802.

[Darroch & Ratcliff, 1972] J.N. Darroch, J. N. & Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. The Annals of Mathematical Statistics 43(5), 1470-1480.

[DeGroot, 1970] DeGroot, M. H. (1970) Optimal Statistical Decisions. New York: McGraw-Hill.

[Della Pietra et al, 1992] Della Pietra, S., Della Pietra, V., Mercer, R. L. & Roukos, S. (1992). Adaptive language modeling using minimum discriminant estimation. Proc. of ICASSP'92, I-633-636.

[Federico 1996] Federico, M. (1996). Bayesian estimation methods for n-gam language model adaptation. Proc. of International Conference on Spoken Language Processing (ICSLP'96), pp.240-243.

[Gauvain & Lee, 1994] Gauvain, J.-L. & Lee, C.-H. (1994). Maximum *a posteriori* estimation for multivariate gaussian mixture observation of Markov chains. IEEE Trans. on Speech and Audio Processing 2, 291-298.

[Hirose, Minematsu, Moriya, 2002] Hirose, K., Minematsu, N. & Moriya, T. (2002). Adaptive training of language models with inter-word co-occurrence for speech recognition. Journal of Information Processing Society of Japan 43(7). (in Japanese)

[Huo & Lee, 1998] Huo, Q. & Lee, C.-H. (1998). On-line adaptive learning of the correlated continuous density hidden Markov models for speech recognition. IEEE Trans. on Speech and Audio Processing 6(4), 386-397.

[Iyer & Ostendorf, 1999] Iyer, R. M. & Ostendorf, M. (1999). Modeling long distance dependence in language: topic mixtures versus dynamic cache models. IEEE Trans. on Speech and Audio Processing 7(1), 30-39.

[Jelinek et al, 1991] Jelinek, F. & Merialdo, B. & Roukos, S. & Strauss, M. (1991). A dynamic LM for speech recognition. Proc. of ARPA workshop on Speech and Natural Language, 293-295.

[Kalai et al, 1999] Kalai, A. & Chen, S. & Blum, A. & Rosenfeld, R. (1999). On-line algorithms for combining language models. Proc. of ICASSP'99,2175-2178.

[Katz87] Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Trans. on Acoustic, Speech and Signal Processing 35(3), 400-401.

[Kawahara et al, 2000] Kawahara, T. *et. al.* (2000). Free software toolkit for Japanese large vocabulary continuous speech recognition. Proc. of International Conference on Spoken Language Processing (ICSLP'2000), Vol.4, 476–479.

[Kawahara & Doshita, 1999] Kawahara, T. & Doshita, S. (1999). Topic independent language model for key-phrase detection and verification. Proc. of ICASSP'99, 685-688.

[Kneser & Steinbiss, 1993] Kneser, R. & Steinbiss, V. (1993). On the dynamic adaptation of stochastic language models. Proc. of ICASSP'93, 586-589.

[Kneser, Peters & Klakow, 1997] Kneser, R, Peters, J. & Klakow, D. (1997). Language model adaptation using dynamic marginals. Proc. of European Conference on Speech Communication and Technology (Eurospeech'97), 1971-1974.

[Kuhn & De Mori, 1990] Kuhn, R. & De Mori, R. (1990). A cache-based natural language model for speech recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence 12(6), 570-583.

[Lasry & Stern, 1984] Lasry, M. J. & Stern, R. M. (1984). A posteriori estimation of correlated jointly Gaussian mean vectors. IEEE Trans. on Pattern Analysis and Machine Intelligence 1(6), 530-535.

[Lau, Rosenfeld & Roukos, 1993] Lau, R., Rosenfeld, R. & Roukos, S. (1993). Trigger-based language models: a maximum entropy approach. Proc. of ICASSP'93, Vol. 2, 45-48.

[Leggetter & Woodland, 1995] Leggetter, C. J. & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation oc continuous density hidden Markov models. Computer, Speech and Language 9, 171-185.

[Matsunaga et. al, 1992] Matsunaga, S., Yamada, T. & Shikano, K. Task adaptation in stochastic language models for continuous speech recognition. Proc. of ICASSP'92, 165-168.

[Masataki et al, 1997] Masataki, H., Sagisaka, Y., Hisaki K. & Kawahara, T. (1997). Task adaptation using MAP estimation in n-gram language modeling. Proc. of ICASSP'97, Vol. 2, 783-786.

[Moriya et al, 2001] Moriya, T., Hirose, K., Minematsu, N. & Jiang, H. (2001). Enhanced MAP adaptation of n-gram language models using indirect correlation of distant words. Proc. of '2001 IEEE workshop on Automatic Speech Recognition and Understanding (ASRU'2001).

[Rao, Monkowski & Roukos, 1995] Rao, P.S., Monkowski, M. D. & Roukos, S. (1995). Language model adaptation via minimum discrimination information. Proc. of ICASSP'95, 161-164.

[Rao, Dharanipragada & Roukos, 1997] Rao, P. S., Dharanipragada, S. & Roukos, S. (1997). MDI adaptation of language models across corpora. Proc. of European Conference on Speech Communication and Technology (Eurospeech'97), 1979-1982.

[Rosenfeld, 1996] Rosenfeld, R. (1996). A maximum entropy approach to adaptive statistical language modeling. Computer, Speech and Language 10(3), 187–228.

[Sasaki, 2000] Sasaki, K. (2000). Language model adaptation by using inter-word correlation for speech recognition, *Master thesis*, the University of Tokyo, Tokyo, Japan. (in Japanese)

[Sasaki, Jiang & Hirose, 2000] Sasaki, K., Jiang, H. & Hirose, K. (2000). Rapid adaptation of n-gram language models using inter-word correlation for speech recognition. Proc. of International Conference on Spoken Language Processing (ICSLP'2000).

	No. article	No. paragraph	No. sentence	No. words
TI Training set	5000	20,274	50,983	1,214,916
Adaptation set A	–	–	100	3192
Adaptation set B	–	–	300	9614
Adaptation set C	–	–	500	16,084
Adaptation set D	–	–	750	24,212
Adaptation set E	–	–	1,000	32,284
Evaluation set	–	–	100	3178

Table 1: Text corpus used in the first experiment on topic adaptation within domain.

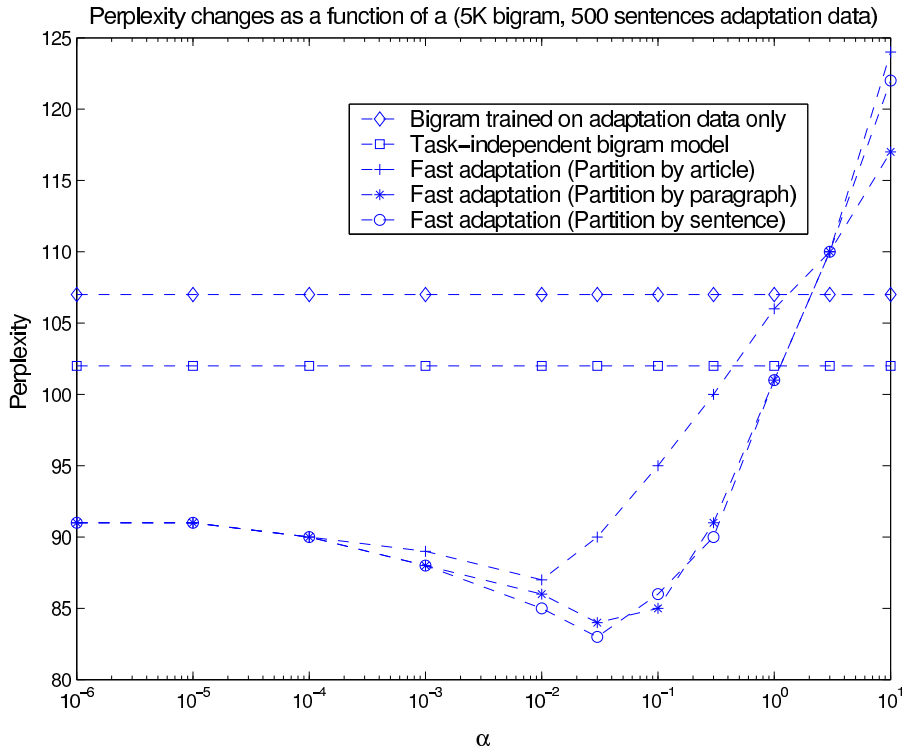

Figure 1: Perplexity changes as a function of α for different partition methods (in sentence, paragraph, or article) in case of using adaptation set C (500 sentences) to adapt the baseline bi-gram model, where we fix $\beta = 1$ and $\lambda = 0$.

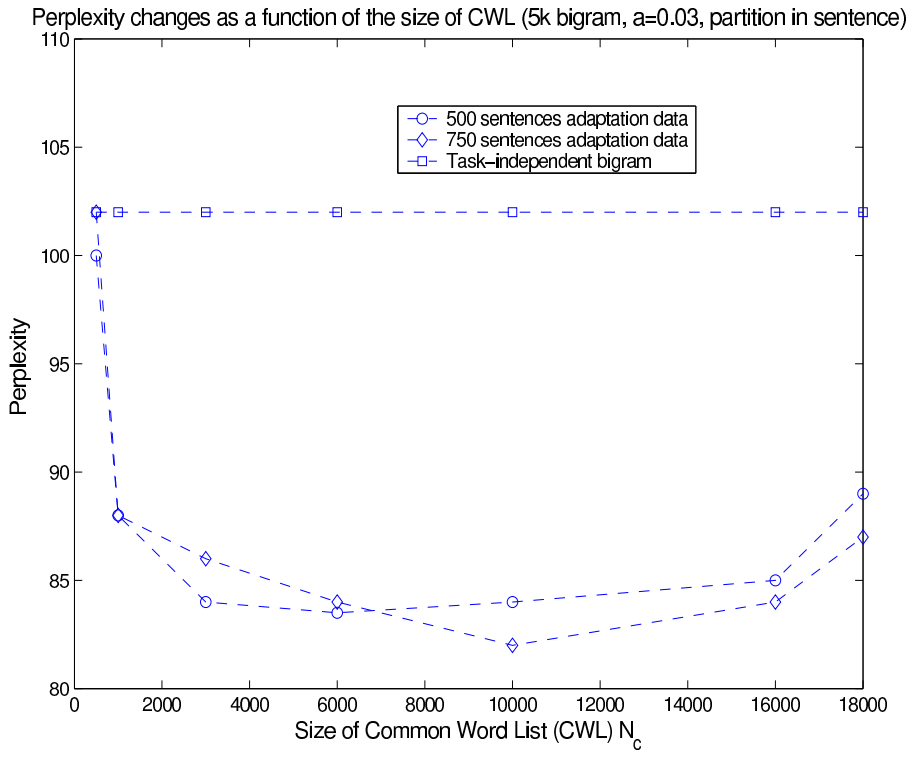

Figure 2: Perplexity changes as a function of CWL size N_c when using adaptation set C (500 sentences) and adaptation set D (750 sentences). ($\alpha = 0.03$ and partition in sentence)

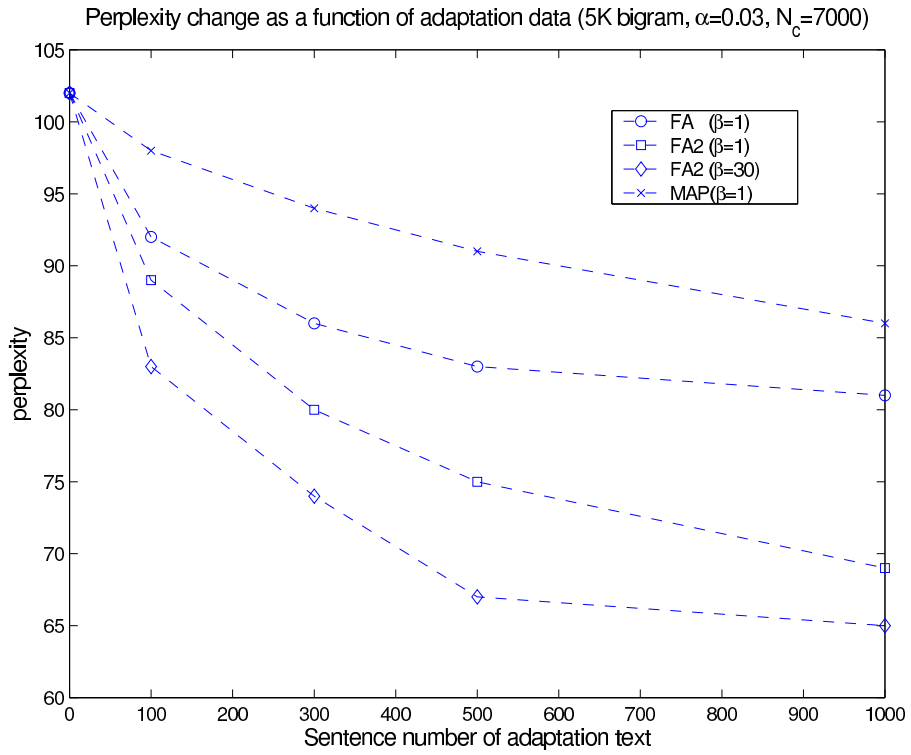

Figure 3: Comparative perplexity reduction results of FA, FA2 and MAP as a function of various amount of adaptation data, where we set $\alpha = 0.03$, $\lambda = 0.03$, and $N_c = 7000$.

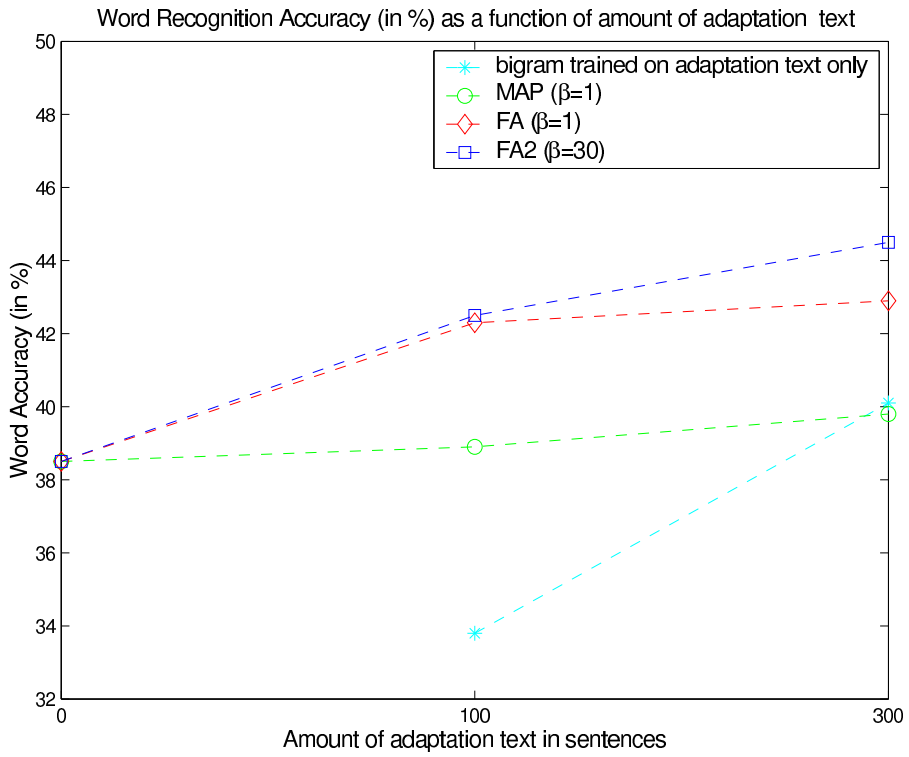

Figure 4: Comparative speech recognition results based on different language model adaptation methods.

	Domain	No. sentence	No. words	No. diff words
TI Training set	newspaper	2,438,662	58,290,111	200,380
Adaptation set 1	*Peter Pan*	133	2,156	639
Adaptation set 2	*Peter Pan*	529	8,931	1,738
Adaptation set 3	*Peter Pan*	699	11,766	2,201
Evaluation set	*Peter Pan*	107	1,709	570

Table 2: Text corpus used in the second experiment on cross-domain adaptation.

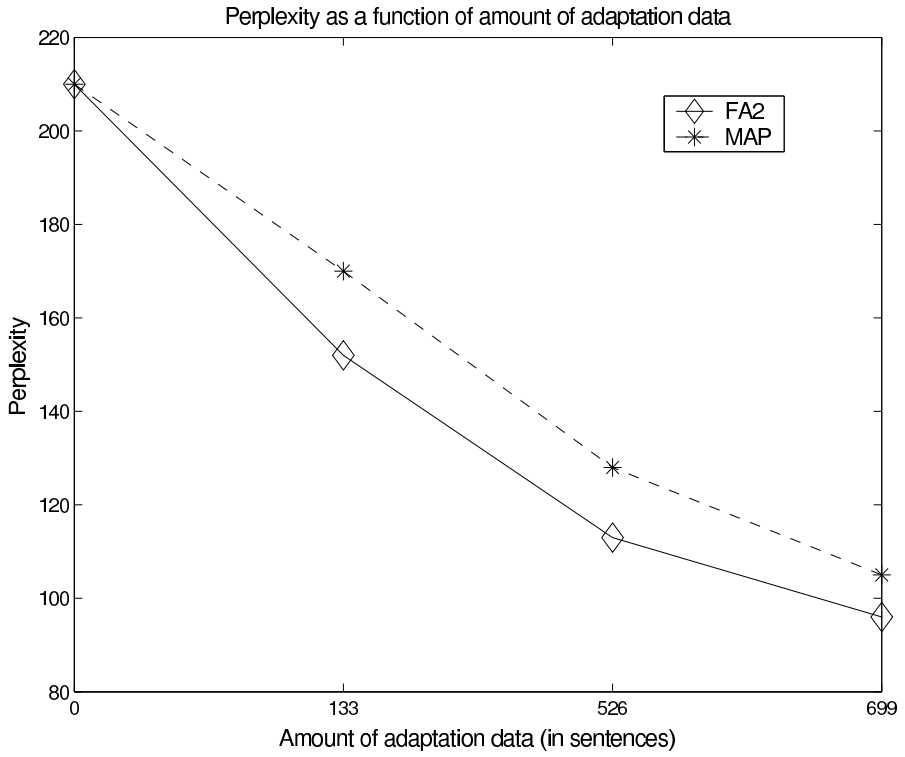

Figure 5: We plot perplexity as a function of amount of adaptation data (in sentences) for the standard MAP adaptation and the proposed fast adaptation method FA2 in cross-domain LM adaptation. In MAP, we choose $\beta = 2500$. In FA2, we choose $\alpha = 1e10$, $\beta = 2500$, $\lambda = 1e7$, $N_c = 8000$.

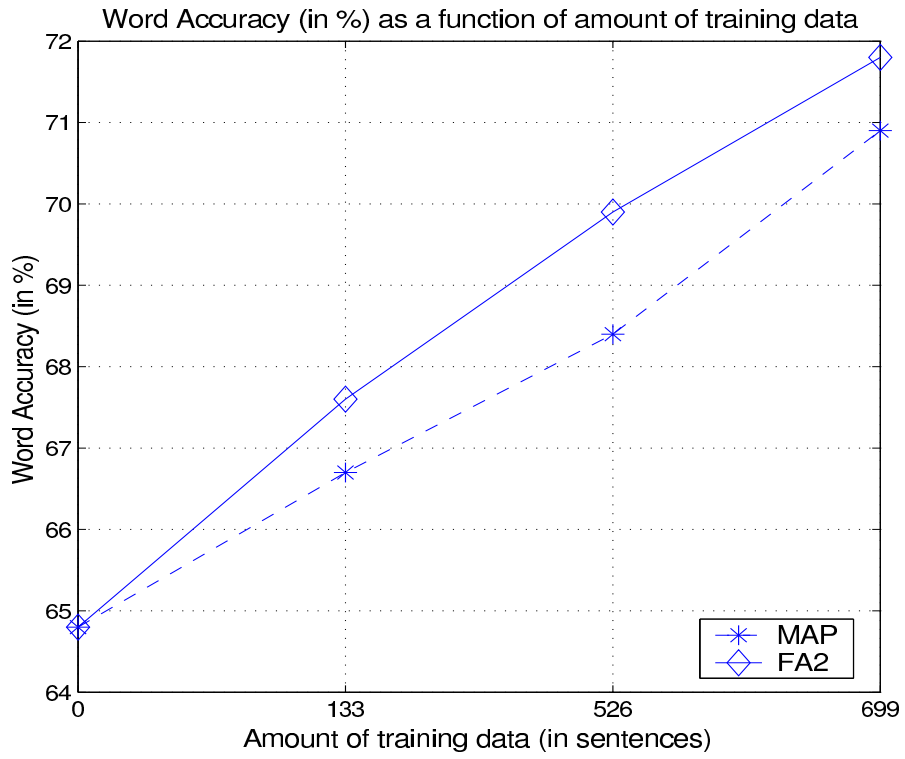

Figure 6: Speech recognition performance (word accuracy in %) as a function of amount of adaptation data (in sentences) for the standard MAP adaptation and the proposed fast adaptation method FA2 in cross-domain LM adaptation. In MAP, we choose $\beta = 2500$. In FA2, we choose $\alpha = 1e10$, $\beta = 2500$, $\lambda = 1e7$, $N_c = 8000$.

List of Figures

List of Tables