

## 複数のバックオフ $N$ -gram を動的補間する言語モデルの高精度化

高橋 力矢<sup>†</sup> 峯松 信明<sup>†</sup> 広瀬 啓吉<sup>††</sup>

<sup>†</sup> 東京大学大学院情報理工学系研究科

<sup>††</sup> 東京大学大学院新領域創成科学研究科

E-mail: †{rikiya,mine,hirose}@gavo.t.u-tokyo.ac.jp

**あらまし** 本稿では複数の  $N$ -gram モデルを動的補間する言語モデルとその高精度化手法について述べる。複数の  $N$ -gram モデルを動的補間する言語モデルは二つの要素技術から成り立つ。それは学習テキスト内のドメイン毎に最適化されたサブ言語モデル群と、各言語モデルの動的補間比を認識仮説や履歴から推定するための手法である。モデル学習に関しては、繰り返しアルゴリズムを用いるが、バックオフ  $N$ -gram を構築していく際に、通常の学習アルゴリズムではオープンデータに対する精度が確保出来ない。そこで本稿では新たにスムージングとの組み合わせを考慮した学習手順を提案している。補間比推定に関しては、Dirichlet 分布を事前分布としたベイズ推定を用いる。モデルの評価は補正パープレキシティ及び大語彙音声認識における認識率の双方から行った。その結果、trigram モデルを動的補間した場合でオープンデータに対して約 5% の補正パープレキシティ低下を得た。認識率に関しては従来の trigram モデルより若干低下または上昇した。

**キーワード** 大語彙音声認識, 複数の言語モデル,  $N$ -gram, 動的補間, バックオフスムージング

## Performance Improvement of the Language Model Using Adaptive Interpolation of Multiple Back-off $N$ -gram Models

Rikiya TAKAHASHI<sup>†</sup>, Nobuaki MINEMATSU<sup>†</sup>, and Keikichi HIROSE<sup>††</sup>

<sup>†</sup> Graduate School of Information Science and Technology, University of Tokyo

<sup>††</sup> Graduate School of Frontier Sciences, University of Tokyo

E-mail: †{rikiya,mine,hirose}@gavo.t.u-tokyo.ac.jp

**Abstract** This paper proposes a new language modeling technique implemented as a multi-mixture model by adaptively interpolating sub-models. Multi-mixture language models (LMs) are built on two components. One is a group of sub-models optimized for domains found in training data and the other is a scheme for context-adaptive and efficient interpolation of the sub-models based on history in recognition hypotheses. In this work, for training sub-models, iterative algorithms of training multiple  $N$ -gram models are adopted. In our case, however, models trained by traditional EM algorithms could not outperform in open data. Thus, a new scheme connecting training algorithms and back-off smoothing of the LMs is also proposed. As for interpolating the sub-models, Bayesian inference using Dirichlet distribution for the prior distribution is adopted. Evaluation of the proposed technique is carried out in two viewpoints; perplexity reduction and large-vocabulary continuous speech recognition (LVCSR) performance improvement. In the first, the proposed LM achieves about 5% relative reduction of adjusted perplexity in an open data set compared to traditional trigram models. In the second, however, the recognition performance got little worse or only slightly improved.

**Key words** LVCSR, multiple LMs,  $N$ -gram, dynamic interpolatoion, back-off smoothing

## 1. はじめに

大語彙音声認識は認識対象のドメインに適応化された言語モデルを用いれば認識精度が上昇する。しかしながら多くの場合、認識対象のドメインは未知であり、認識結果の履歴を参照して推定する必要がある。ましてドメインが時間とともに変動する場合を考えると、ドメインの推定は一文程度の短い履歴単語列で推定しなくてはならない。キャッシュモデルやトリガーモデル [1], [2] はこのような履歴に対する動的適応を効率的に実現してきた言語モデルである。これらのアプローチは履歴内の単語と関連のある単語の出現確率を上昇させるものである。一方、Probabilistic Latent Semantic Indexing (PLSI) や Latent Dirichlet Allocation (LDA) [3], [4], はこれを発展させて、ドメイン毎に複数用意された unigram の混合モデルを用いて単語出現確率を与えた。Hofmann らは trigram を用いる際には  $N$ -gram 確率を次式で変形する unigram rescaling [5] を用いていた。

$$P(w_n|w_{n-N+1}^{n-1}) \propto \frac{P(w_n|h_n)}{P(w_n)} P(w_n|w_{n-N+1}^{n-1}) \quad (1)$$

但し  $h_n$  は  $N$  よりも長い文脈である。PLSI により unigram rescale された trigram モデルは 17% のパープレキシティ低下を達成していたが、この手法は次単語を予測する度に出現確率を正規化する必要がある。仮説ベースの大語彙音声認識においてはこの計算時間がボトルネックとなる。

それを受けて筆者らは [6] で複数のバックオフ  $N$ -gram 言語モデルを直接動的補間する方法を提案したが、trigram モデルの補間においてオープンデータに対する性能向上が見られなかった。そこで本稿では、サブ言語モデルの性質に着目したバックオフスムージングとの組合せを用いてオープンデータに対しても有効なモデリングを提案する。

## 2. 提案モデル

提案される言語モデルは、 $N$ -gram サブモデルの混合モデルとして表現される。つまり長さ  $L$  の履歴単語列に依存した単語  $w_n$  の生起確率は式 (2) で表され、式 (3) で近似される。

$$P(w_n|w_{n-L}^{n-1}) = \sum_{i=1}^m P(w_n|w_{n-L}^{n-1}, z_i) P(z_i|w_{n-L}^{n-1}) \quad (2)$$

$$\approx \sum_{i=1}^m P(w_n|w_{n-N+1}^{n-1}, z_i) P(z_i|w_{n-L}^{n-1}) \quad (3)$$

但し、 $z_i$  はドメイン、 $m$  は  $N$ -gram サブモデルの混合数 (異なるドメインの数) である。式 (3) は  $L < N$  の場

合にも有効であるが、この場合静的な補間を意味する。 $P(w_n|w_{n-N+1}^{n-1}, z_i)$  の値は次節に示す学習アルゴリズムによって導出されるが、 $P(z_i|w_{n-L}^{n-1})$  の値は学習アルゴリズムでは計算されず評価・認識時に  $P(w_n|w_{n-N+1}^{n-1}, z_i)$  の値を利用して算出する。この理由であるが、本来  $P(z_i|w_{n-L}^{n-1})$  の記録には語彙数  $V$  に対して  $O(V^L)$  の記憶領域が必要であり、現在のところ実装不可能なためである。加えて固定された値を用いると、バックオフスムージングのような手法以外には、学習データに登場しなかった長さ  $L$  の単語に対して確率値を割り当てることが出来ない。

### 2.1 学習アルゴリズム

言語モデルの学習は、通常学習テキストの尤度を最大化することで行われる (最尤推定) が、式 (3) をモデリングした場合、通常最尤推定では明確なアルゴリズムが導出出来ない。そこで本研究では式 (3) に関連した別の尤度を最大化することで学習を行う。それぞれの文  $s$  に対して  $P(z_i|s)$  を定めると、次式に示す対数尤度  $\mathcal{L}$  が最大化される。

$$\mathcal{L} = \sum_{s \in S} \sum_{h, w \in s} n(s, hw) \log P(w|h, s) \quad (4)$$

$$= \sum_{s \in S} \sum_{h, w \in s} n(s, hw) \log \sum_{i=1}^m P(w|h, z_i, s) P(z_i|s) \quad (5)$$

但し、 $S$  は学習テキストにおける文の集合、 $w$  は単語、 $h$  は  $N-1$  単語からなる履歴単語列である。今回の場合、 $N$ -gram モデルの単語生起確率は全ての  $z_i$  に対して必要であるため、 $P(w|h, z_i, s)$  は近似的に  $P(w|h, z_i)$  と表される。よって式 (6) が得られる。

$$\mathcal{L} = \sum_{s \in S} \sum_{h, w \in s} n(s, hw) \log \sum_{i=1}^m P(w|h, z_i) P(z_i|s) \quad (6)$$

式 (3) に基づく尤度と違い、次に示す尤度を最大化するのは容易である。具体的には、事後確率  $P(z_i|s, hw)$  を導入し、ラグランジュの未定乗数法を用いれば良い。この確率値は次式で表される。

$$P(z_i|s, hw) = \frac{P(w|h, z_i) P(z_i|s)}{\sum_{i=1}^m P(w|h, z_i) P(z_i|s)} \quad (7)$$

最後に、得られた繰り返し学習アルゴリズムを図 1 に示す。但し、 $W_h$  は  $h$  の後に出現する異なり単語の空間である。このアルゴリズムにおいては、ドメインの推定は  $s$  毎にされると仮定している。次節以降に示すように、 $L$  は経験的なパラメータであり、一文の中の単語数と関連している。そのため、著者らは  $L$  に関するこのような仮定は問題無いと考えている。

### 2.2 提案手法における Back-off スムージング

一般に、学習テキストにない単語列がテストセットに出現した場合、バックオフスムージングを介して低次の

**E-step:**事後確率  $P(z_i|s, hw)$  を算出

$$P(z_i|s, hw) = \frac{P(w|h, z_i)P(z_i|s)}{\sum_{i=1}^m P(w|h, z_i)P(z_i|s)} \quad (8)$$

**M-step:** $P(z_i|s, hw)$  から  $P(w|h, z_i)$ ,  $P(z_i|s)$  を算出

$$P(w|h, z_i) = \frac{\sum_{s \in S} n(s, hw)P(z_i|s, hw)}{\sum_{w \in W_h} \sum_{s \in S} n(s, hw)P(z_i|s, hw)} \quad (9)$$

$$P(z_i|s) = \frac{\sum_{hw \in s} n(s, hw)P(z_i|s, hw)}{\sum_{i=1}^m \sum_{hw \in s} n(s, hw)P(z_i|s, hw)} \quad (10)$$

図 1 複数の  $N$ -gram サブモデルに対する学習アルゴリズム

言語モデルが要求される。通常、 $N$ -gram モデルの単語出現頻度が与えられた場合、 $(N-1)$ -gram モデルにおける単語出現頻度は  $N$ -gram モデルにおけるそれらの和として表せる。しかし本稿の混合モデルでは、このようなモデル導出法を用いるとオープンデータに対する性能が発揮出来ないことがわかっている [6]。その理由として、提案学習アルゴリズムを用いて学習を行った後は、この導出法に基づく低次モデル (つまり  $(N-1)$ -gram 以降) を算出した場合、それらの低次モデルはドメインにあまり依存せず均質なモデル群になるためである。

しかし低次モデルを高次のモデルとは独立に学習した場合、 $N$ -gram モデルに対するドメインの集合と、 $(N-1)$ -gram モデルに対するそれらとは、対応がとれないことになる。この問題を解決する方法として、低次言語モデルの性質に着目した学習アルゴリズムについて記述する。このアルゴリズムの事後処理としてのバックオススムージングについても記述する。

**2.2.1 低次モデルを先鋭化する学習方法**

低次モデル群を学習する際には、それらの持つドメインと高次モデル群のそれとが対応したままである必要がある。簡単のため、ここでは bigram 混合モデルと unigram 混合モデルとのドメイン対応を考える。提案学習アルゴリズムは  $P(z_i|s, hw)$  の分布を先鋭化するが、これは  $w$  の前に  $h$  が出現したという条件下での  $P(z_i|s, w)$  の値である。従って、unigram 混合モデルを学習した際の  $P(z_i|s, w)$  の収束値を bigram 混合モデルを学習する際の  $P(z_i|s, hw)$  の初期値として採択すれば、繰り返したアルゴリズムにおいてドメインの対応はずれないと考えられる。同じ事が bigram 混合と trigram 混合の間にも言える。図 2 に複数の  $N$  に対する  $N$ -gram サブモデル群を学習するアルゴリズムを示す。なお、図 2 で括弧書きで書かれている内容は 2.2.2 節に示した学習手順であり、本節の手順とは別である (スペース省略のためまとめた)。

**2.2.2 高次モデルを統合していく学習方法**

しかし、EM アルゴリズムにおける局所解を極力避けるためには、低次の言語モデルからよりも、最も重要な最高次の言語モデルから学習すべきである。図 2 は  $N = 1$  からスタートしているが、trigram ならば  $N = 3$  から始めて  $N \leftarrow N - 1$  としていっても学習できる。この場合はドメインが文ごとに振られているため、単純な出現頻度の和で低次言語モデルを算出する場合とは違うモデルが得られる。ここに示した学習方法についても図 2 内に括弧書きで示した。

**2.2.3 ディスカウンティング**

バックオススムージングのためのディスカウンティングには、Witten & Bell [7] スムージングを採用する。式 (9) において、

$\sum_{s \in S} n(s, hw)P(z_i|s, hw)$  を  $C(hw|z_i)$ 、つまり“ドメイン  $z_i$  における  $hw$  の出現頻度”とみなす。Witten & Bell スムージングを用いる際には、 $r(h|z_i)$ 、つまり“ドメイン  $z_i$  における単語列  $h$  の後の異なり単語数”も必要であるが、それは次の二式により求まる。

$$Q(hw|z_i) = \begin{cases} 1 & C(hw|z_i) \geq 1 \\ C(hw|z_i) & C(hw|z_i) < 1 \end{cases} \quad (11)$$

$$r(h|z_i) = \sum_{w \in W_h} Q(hw|z_i) \quad (12)$$

このとき、スムージング手順は次式でまとめられる。

$$P(w|h, z_i) = \begin{cases} \frac{C(hw|z_i)}{C(h|z_i) + r(h|z_i)} & C(hw|z_i) > 0 \\ \frac{r(h|z_i)P(w|h', z_i)}{C(h|z_i) + r(h|z_i)} & C(hw|z_i) = 0 \end{cases} \quad (13)$$

但し  $h'$  は  $h$  内の単語の中で、最も古い単語 (つまり  $w$  から離れた単語) を取り除いた単語列である。

**3. サブ言語モデルの動的補間**

図 1, 2 における学習アルゴリズムは  $P(w|h, z_i)$  を

- (1)  $N = 1$  [ $N = 3$ ] とする。
- (2) 図 1 のアルゴリズムを用いて、ドメイン  $z_i$  に対して  $N$ -gram モデル群を学習する。
- (3) 各々の  $z_i$  に対して、 $N$ -gram モデル群を学習した際の  $P(z_i|s, hw)$  の収束値を  $(N+1)$ -gram [( $N-1$ )-gram] モデル群の  $P(z_i|s, hw)$  の初期値とする。
- (4)  $N \leftarrow N + 1$  [ $N \leftarrow N - 1$ ] として step 2 へ。
- (5) それぞれのサブ言語モデルに対して確率値のディスカウントを行う。

図 2 バックオフ言語モデルのための学習手順

与える。2. 節に記載したように、 $P(w|h, z_i)$  を用いて  $P(z_i|w_{n-L}^{n-1})$  を得るアルゴリズムが必要である。 $w_{n-L}^{n-1}$  は学習テキストのものではなく認識仮説における単語列であることに注意する必要がある。 $P(z_i|w_{n-L}^{n-1})$  の予測には現在のところ、ベイズ推定を用いている。これは文献 [6], [8] に記載されたものと同一である。次の二式を収束するまで交互に繰り返し、パラメータ  $\overline{N_{n,i}}$  を算出する。

$$Q(y_i^{n-k} = 1|w_{n-L}^{n-1}) \propto \exp[\Psi(\phi_0 + \overline{N_{n,i}})] \\ \times P(w_{n-k}|w_{n-k-N+1}^{n-k-1}, z_i) \quad (14)$$

$$\overline{N_{n,i}} = \sum_{k=1}^L Q(y_i^{n-k} = 1|w_{n-L}^{n-1}) \quad (15)$$

但し、 $\Psi$  は対数ガンマ関数の第一微分である digamma 関数である。最終的には  $E[\lambda_{n,i}]$  は次式で与えられる。

$$E[P(z_i|w_{n-L}^{n-1})] = \frac{\phi_0 + \overline{N_{n,i}}}{\sum_{i=1}^m (\phi_0 + \overline{N_{n,i}})} \quad (16)$$

超パラメータ  $\phi$  の値は、Jeffreys-Perks law に従い 0.5 と定める。

#### 4. 評価実験-1

提案モデルをパープレキシティで評価した。毎日新聞 8 年分 (1991 年度から 1998 年度まで) コーパスを準備し、そのうち 7 年分を学習テキスト、1 年分を評価テキストとして採用した。全ての実験において、語彙数は新聞 8 年分に対する上位 2 万語とした。

##### 4.1 提案モデル及びベース言語モデルの学習

1991 年から 1997 年までの新聞テキストを用いて、図 1, 2 の手順に従い bigram 混合モデル及び trigram 混合モデルを学習した上で、式 (11)(13) を用いてスムージングした。なお各モデルに対し、2.2.1 に示した低次モデルから学習したもの (タイプ A) と、2.2.2 に示した高次モデルから学習したもの (タイプ B) との二種類ずつを構築した。

比較のためのベース言語モデルとして、通常のバックオフ bigram モデル及びバックオフ trigram モデルを同一の学習テキストより構築した。そのスムージングにはやはり Witten-Bell スムージングを用いた。

##### 4.2 テストセットの選定

テストセットは 1998 年度テキストのサブセットとした。提案モデルは、入力されるドメインが変化してもそれに動的に追従する。この追従精度を評価するために、1998 年度テキストから話題が適宜変わるようなテキストを生成してこれを評価テキストとした。毎日新聞コーパスには 17 種類の手動分類された記事ラベルがある。1998 年度テキストをこのラベルに基づいて 17 種類のサブテキスト  $T_k; \{1 \leq k \leq 17\}$  に分割した。図 3 に示した手順

- (1) それぞれのサブテキストに対し読み取り位置をその先頭に置く。
- (2) 無作為に一つのサブテキスト  $T_k$  を選び、読み取り位置から連続する  $X$  文を採取する (読み取り位置は  $X$  文進む)。これらの  $X$  文をテストセットに加える。
- (3)  $T_k$  の読み取り位置を更に  $Y$  文を進める。
- (4) テストセットとして 1000 文が得られるまで step 2 に飛び処理を繰り返す。

図 3 テストセットの生成手順

表 1 ドメイン変遷の 3 種類のレベル

Name	Instances of X and Y
Slow	$1 \leq X \leq 10, 1 \leq Y \leq 10$
Fast	$1 \leq X \leq 3, 1 \leq Y \leq 10$
VeryFast	$X = 1, 1 \leq Y \leq 100$

に基づき、これらのサブテキストからいくつかのテストセットを生成した。図 3 で、 $X, Y$  は表 1 に示された範囲の変数。表 1 に示すように、ドメイン変遷速度が異なる 3 種類のテキストを用意した。

これらとは別に、ATR 503 音素バランス文もテストセットに加えた。ATR503 文内の各文は新聞、雑誌、小説、手紙、教科書など様々な範疇のテキストから取り出されているため、話題変遷がとても早く、かつ各々のドメインが新聞評価テキストに比べて学習テキストから「より」離れている、と解釈できる。

##### 4.3 実験結果

補正パープレキシティ ( $APP$ ) を評価指標として用いた。表 2 にベース言語モデルによる各評価テキストの  $APP$ 、総単語数  $WC$ 、未知語率  $OOV$  を示した。図 4 には各評価テキストに関して、ベース言語モデルからの  $APP$  の変化率を示した。図 4 ではモデルの混合数は 8 に固定してある。タイプ A に関しては、bigram 混合モデル、trigram 混合モデルの双方とも通常の  $N$ -gram よりパープレキシティが低下しており、どのテキストでも  $L = 10$  付近で極小となっている。これは [6] において trigram でパープレキシティ低下が得られなかった結果から進歩している。新聞評価テキストに関して言えば、ど

表 2 ベース言語モデルによるテストセットの  $APP_{base}$ ,  $WC$ ,  $OOV$

text	$APP$ by bigram	$APP$ by trigram	$WC$	$OOV$
Slow	88.52	56.82	25953	3.83%
Fast	85.65	54.48	26953	3.62%
VeryFast	85.40	55.41	26551	3.58%
ATR503	106.859	77.90	9296	6.57%

のモデルを用いても最適性能は履歴 10 単語から 20 単語程度で得られ、 $L$  がそれ以上大きくなると適応が静的となり言語モデリングの意味を失う。テストセットにおけるドメイン変遷速度の違いに着目した場合、最小の  $APP$  を与える  $L$  の値は速度が遅くなるほど大きくなる。つまり話題追従するかしないかが結果に現れており、これは明らかに提案モデルが動的適応していることを示唆するものである。

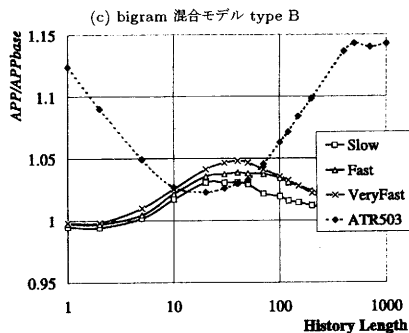
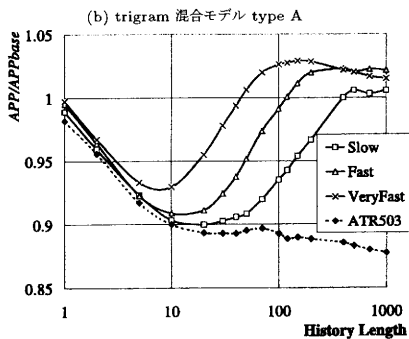
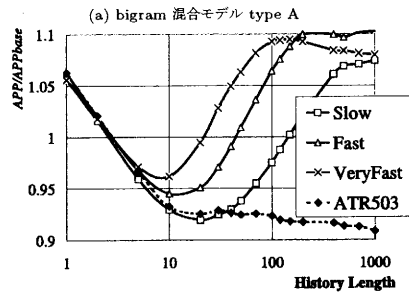
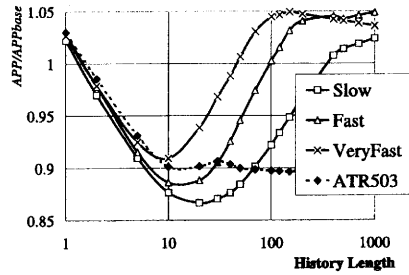
タイプ B に関しては、trigram の場合、全てのオープンデータ評価テキストに関してパープレキシティが低下しなかった。ここには記載していないが、予備実験でタイプ B はクロズドデータに対してはタイプ A よりも若干精度が高いという結果が出ている。これは、本稿のはじめに述べた学習アルゴリズムに注意することでドメインの対応がとれるという予測が否定される結果となっている。これについてはまとめの項で改めて詳しく述べる。

以降、タイプ A に関してのみ実験を行った。提案モデルは、学習テキストからドメインが特に離れた ATR503 文においても、通常の  $N$ -gram より精度が高い。しかしながら、 $L$  の値が増えれば増えるほどパープレキシティが単調に低下する点で新聞評価テキストとは異なる。これは、学習テキストから離れすぎたドメインの入力においては、話題追従ができず動的適応が出来ないこと、その結果静的な適応が効いてくることを意味している。

以上の実験はモデルの混合数を 8 に固定した場合で行ったが、 $m$  の値を変化させた場合のパープレキシティの変化を調べた。今回は図 4 を踏まえて  $L = 10$  に固定した。図 5 は提案モデルのうち、bigram 混合モデルにおける  $m$  と  $APP/APP_{base}$  との相関を示している。この実験においても、ATR503 文における結果は新聞評価テキストと異なっている。最小の  $APP$  は新聞テキストでは  $m = 8$  で観察されたが、ATR503 文ではそれが  $m = 4$  付近となっている。ATR503 文の持つドメインは新聞のそれと異なるため、これは自然な結果である。文献 [5] では、最適性能を与える unigram 混合モデルの混合数は 100 を越えており、これは我々の結果と大きく異なる。通常、bigram モデルや trigram モデルのパープレキシティは unigram より低くなるため、もともと最適な分割数が低いと考えられる。そのため、unigram モデルの場合は、分割数を多くしても過学習になりにくい。現在のところ混合モデルにおける最適混合数の違いはこれに起因すると我々は考えている。

## 5. 評価実験-2

今回は大語音音声認識における提案モデルの性能を評価する。認識エンジンには 2-パス構成の JULIUS を用い



(d) trigram 混合モデル type B  
図 4  $APP$  と履歴長との関係

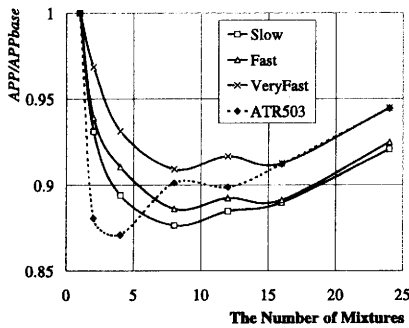


図5 APF と bigram 混合数との関係

表3 ベース言語モデル及び提案モデルによる音声認識精度

speech data	model	word correct rate	word accuracy
1998 News	baseline	88.21%	85.00%
1998 News	proposed( $m = 8$ )	87.62%	84.43%
JNAS-PB	baseline	85.64%	81.99%
JNAS-PB	proposed( $m = 4$ )	85.70%	82.10%

た。JULIUSにおいて言語モデルは、1stパスには順方向 bigram を用い、2ndパスには逆向き trigram を用いている。提案する混合モデルはそのままでは JULIUS に実装できないため、内部モジュールを書き換えて2ndパスに提案モデルを実装した。用いたモデルは混合数4及び8の逆向き trigram 混合モデル(学習方法タイプA)である。音響モデルは状態数約2000の tied-state triphone を用いた。テスト発声として、二種類の音声を用意した。一つ目は1998年度評価テキストの表1における“VeryFast”のサブセットで男性話者一名による419文である。二つ目はJNAS-PB databaseのサブセット1800文で、これらの文は前節実験のATR503文と深い関係がある。テストセットの発声は18名の男性及び18名の女性話者によって行われている。表3にベース言語モデル及び提案モデルによる認識精度を示した。提案モデルは前節のパープレキシティ評価ではいくつかの改善が見られたが、認識率は少しばかり低下している。これらは現在のところ、認識仮説内の誤った文への適応のためと考えられる。また文献[5]においても認識率への寄与が小さかったことを踏まえると、このようなドメイン推定によって同定される単語はもともと音響的に他単語と区別しやすく、認識精度の改善に寄与しにくいという可能性も考えられる。

## 6. まとめと今後の展望

$N$ -gram サブモデルの混合モデルを動的補間する言語モデルを提案した。本稿では新たにサブモデルの学習方

法に関して、全ての  $N$  に対してモデル学習するための手順を示した。認識時のドメイン推定には、モデル混合比の確率分布に関し、ディレク分布を事前分布としたベイズ推定を行った。評価実験では、提案手法により従来の trigram モデルから補正パープレキシティが約5%低下した。しかしながら音声認識における認識精度はわずかに低下したか従来の  $N$ -gram と変わらないかのどちらかであった。この原因として誤った仮説に対する不適切な適応が考えられる。

また低次モデルから学習した混合モデルと高次モデルから学習した混合モデルとでオープンデータに対する精度が異なったことから、unigram サブモデルのドメインと trigram サブモデルのドメインとが本質的に異なり、より正確な対応が不可能であることが示唆された。オープンデータに対する trigram カバー率は新聞7年分学習しても、88%程度であることを踏まえると、バックオフ  $N$ -gram を複数用意して補間するのではなく、今後はPLSIやLDAに代表される unigram の混合モデルを元に trigram を補間する言語モデリングを試みていく。また誤認識仮説への適応を避けるため、予測単語の近傍単語列に対して頑健な確率値を出すクラス言語モデルとの組合せを考案していく。

## 文 献

- [1] R. Kuhn and R. De. Mori, “A cache-based natural language model for speech recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 570–583, 1990.
- [2] R. Rosenfeld, “A maximum entropy approach to adaptive statistical language modeling,” *Computer Speech and Language*, vol. 10, no. 5, 1996.
- [3] T. Hofmann, “Probabilistic Latent Semantic Indexing,” in *Proc. of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, Berkeley, California, August 1999, pp. 50–57.
- [4] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, vol. 3, no. 5, pp. 993–1022, 2003.
- [5] D. Gildea and T. Hofmann, “Topic-based language models using em,” in *Proc. of the 6th European Conference on Speech Communication and Technology (EUROSPEECH)*, 1999, pp. 2167–2170.
- [6] 高橋 力矢, 峯松 信明, and 広瀬 啓吉, “文脈適応による複数  $n$ -gram の動的補間を用いた言語モデル,” in *情報学研報*, 2003, 2003-SLP-46-7, pp. 37–42.
- [7] I. H. Witten and T. C. Bell, “The zero frequency problem: Estimating the probabilities of novel events in adaptive text compression,” *IEEE Transactions on Information Theory*, vol. 37, no. 4, pp. 1085–1093, 1991.
- [8] 三品 拓也 and 山本 幹雄, “確率的 LSA に基づく  $n$ -gram モデルの変分ベイズ学習を利用した文脈適応化,” in *信学技報*, 2002, NLC-2002-73, pp. 13–18.