音声言語情報処理 45−15
（２００３．２．８）

単語クラスタリングによる確率的言語モデルの分野適応

森 信介　伊東 伸泰　西村 雅史
日本アイ・ビー・エム東京基礎研究所
〒 242-8502 神奈川県大和市下鶴間 1623-14
{mori,iton,nisimura}@trl.ibm.com

あ ら ま し
音声認識などに利用される確率的言語モデルの構築は、対象分野のコーパスを大量に必要とする。対象分野によっては、十分なコーパスが存在せず、これが認識精度向上の障害となっている。本論文では、対象とは分野が異なるが大量に存在するコーパスを用いることで、対象分野のコーパスのみから推定された言語モデルより高い予測力を持つ言語モデルを構築する方法を提案する。実験では、コーパスとして放送大学の書き起こしと日経新聞の記事を用いて、従来の線形補間による方法と提案手法の予測精度を比較した。パープレキシティーの比較の結果、提案手法によるモデルは線形補間によるモデルに対して 9.71% の改善が確認された。

キーワード　確率的言語モデル　分野適応　クラスタリング　補間　音声認識

Language Model Adaptation Using Word Clustering

Shinsuke MORI, Nobuyasu ITOH, Masafumi NISHIMURA
IBM Research, Tokyo Research Laboratory, IBM Japan, Ltd.
1623-14 Shimotsuruma Yamatoshi Kanagawaken 242-8502 Japan
{mori,iton,nisimura}@trl.ibm.com

Abstract
Building a stochastic language model (LM) for speech recognitions, etc. requires a large corpus of target task. In some tasks no enough large corpus is available and this is an obstacle to achieve a high recognition accuracy. In this paper, we propose a method for building an LM with a higher prediction power using large corpora of different tasks than an LM estimated from a small corpus of a target task. In our experiment, we used transcriptions of air university lectures and articles of *Nikkei* newspaper and compared an existing interpolation-based method and our new method. The result tells us that our method allows 9.71% of perplexity reduction.

Key Words Stochastic Language Modeling, Task Adaptation, Clustering, Interpolation, Speech Recognition

1 はじめに

音声認識などに利用される確率的言語モデルの構築は、対象分野のコーパスを大量に必要とする。対象分野によっては、十分なコーパスが存在せず、これが認識精度向上の障害となっている。

この問題の解決法として、補間を応用することが提案されている [1, 2, 3]。具体的には、対象分野の小量のコーパスから推定された言語モデル (n-gram モデルなど) と新聞などの大規模なコーパスから推定された言語モデル (n-gram モデルなど) を線形補間する。

本論文では、適応する分野の小規模コーパスには出現しない語彙であっても、一般的な大規模コーパスに出現すれば、適応する分野から得られる文脈情報も予測に用いることを可能にする方法を提案する。これは、適応する分野の単語をクラスとみなし、一般的な大規模コーパスに出現する語彙を類似のクラスに割り振ることで実現される。

有効性を実験的に示すために、コーパスとして放送大学の書き起こしと日経新聞の記事を用いて、従来の線形補間による方法と提案手法の予測能力を比較した。パープレキシティーの比較の結果、提案手法によるモデルは線形補間によるモデルに対して 9.71%の改善が確認された。

2 言語モデル

本論文で提案する分野適応手法は、文をある単位の列と見なすあらゆる確率的言語モデルに適用可能である。本論文では、形態素 n-gram モデルへの適用について説明し、その実験結果を示す。この節では、形態素 n-gram モデルと提案する分野適応手法を適用した結果得られるクラス n-gram モデルを説明する。

2.1 形態素 n-gram モデル

形態素 n-gram モデルは、文を形態素の列 ($\boldsymbol{m} = m_1 m_2 \cdots m_h$) と見なし、各形態素を文の先頭から順に予測する。このとき、次式が示すように直前の $k = n-1$ の形態素を条件とする。式が簡便になるように、先頭の形態素の前には文頭記号 (BT) が十分長く存在し、最後の形態素の後にも文区切り記号として (BT) が 1 つ存在する仮定する。

日本語の形態素を全て列挙することはできないので、未知形態素の扱いが避けられない問題となる。この問題に対処するため、未知形態素に対応する特別な記号 (UM) を用意し、既知の形態素以外はこの記号から後述する未知語モデルにより与えられる確率で生成されることとする。未知形態素に対応する特別な記号は、かならずしも唯一である必要はなく、品詞などの情報を用いて区別される複数の記号であってもよい。以下の説明では、各品詞に対して未知形態素に対応する記号 ($\mathtt{UM}_{pos}$) を設ける。

以上に述べた形態素 n-gram モデル M_m による、形態素列 $m_1 m_2 \cdots m_h$ の出現確率は以下の式で表される。ただし $\mathcal{M}_k$ は既知形態素の集合を表し、pos は m_i の品詞を表す。また $m_i = \mathtt{BT}$ $(i \leq 0 \vee i = h+1)$ である。

$$
\begin{aligned}
& M_m(m_1 m_2 \cdots m_h) \\
&\quad = \prod_{i=1}^{h+1} P_m(m_i | m_{i-k} \cdots m_{i-2} m_{i-1}) \\
& P_m(m_i | m_{i-k} \cdots m_{i-2} m_{i-1}) \\
&\quad = \begin{cases} P(m_i | m_{i-k} \cdots m_{i-2} m_{i-1}) & \text{if } m_i \in \mathcal{M}_k \\ P(\mathtt{UM}_{pos} | m_{i-k} \cdots m_{i-2} m_{i-1}) M_{x,pos}(m_i) & \\ \qquad\qquad \text{if } m_i \notin \mathcal{M}_k \end{cases}
\end{aligned}
$$

この式の中の $M_{x,pos}$ は、次項で述べる未知語モデルであり、品詞が pos であることを条件として、引数で与えられる形態素 m_i の文字列 $\boldsymbol{x}$ の生成確率を値とする。

2.2 未知語モデル

未知語モデルは、表記から確率値への写像として定義され、既知形態素以外のあらゆる形態素の表記を 0 より大きい確率で生成し、この確率をすべての表記にわたって合計すると 1 以下になる必要がある。このような条件を満たすモデルとして、文字単位の n-gram モデルがある。文字 n-gram モデルは、形態素 n-gram モデルの形態素を文字と見做すことで定義できる。形態素 n-gram モデルの場合と同様に、文字集合をコーパスに現れる既知文字 ($\mathcal{X}_k$) と現れない未知文字 ($\mathcal{X}_u$) に分類し、未知文字はこれを表す特別な記号 (UX) から生成されるものとする。文字は有限であるから、形態素 n-gram モデルの場合と異なり、各未知文字の生成確率を等確率 ($1/|\mathcal{X}_u|$) とすることができる。

以上に述べた未知語モデル M_x による、文字列 $x_1 x_2 \cdots x_h$ の出現確率は以下の式で表される。ただし $x_i = \mathtt{BT}$ $(i \leq 0 \vee i = h+1)$ である。

$$
\begin{aligned}
& M_x(x_1 x_2 \cdots x_h) \\
&\quad = \prod_{i=1}^{h+1} P_x(x_i | x_{i-k} \cdots x_{i-2} x_{i-1}) \\
& P_x(x_i | x_{i-k} \cdots x_{i-2} x_{i-1}) \\
&\quad = \begin{cases} P(x_i | x_{i-k} \cdots x_{i-2} x_{i-1}) & \text{if } x_i \in \mathcal{X}_k \\ P(\mathtt{UX} | x_{i-k} \cdots x_{i-2} x_{i-1}) \frac{1}{|\mathcal{X}_u|} & \text{if } x_i \notin \mathcal{X}_k \end{cases}
\end{aligned}
$$

2.3 低頻度事象への対処

一般に、n-gram モデルのパラメータ推定には最尤推定が用いられる。しかし、対象とする事象の頻度が低い場合には、推定値の信頼性は低くなるという問題がある。この問題に対処する方法として、補間と呼ばれる方法が用いられる [4]。これは、次の式で表されるように、より信頼性が

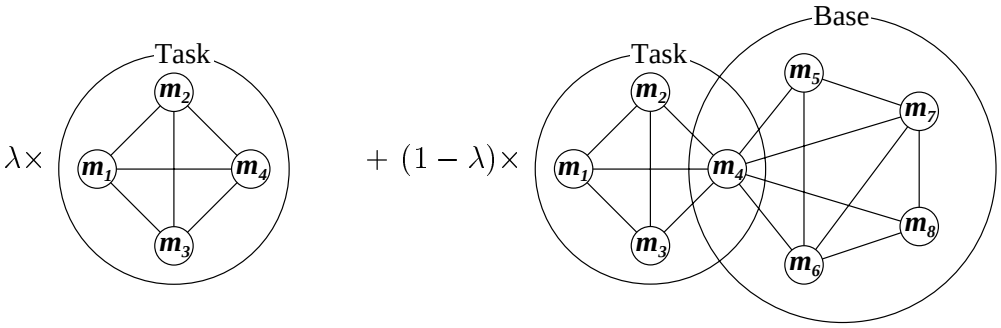


図 1: 線形補間によるによる分野適応

高いことが期待される、より低次の n-gram モデルの確率を一定の割合で足し合わせるという操作を施すことをいう。

- 形態素 n-gram モデル

$$P(m_i|m_{i-k}m_{i-k+1}\cdots m_{i-1}) = \sum_{j=0}^{k}\lambda_j^m P_{MLE}(m_i|m_{i-j}m_{i-j+1}\cdots m_{i-1})$$

ただし $0 \le \lambda_j^m \le 1$, $\sum_{j=0}^{k}\lambda_j^m = 1$

- 文字 n-gram モデル

$$P(x_i|x_{i-k}x_{i-k+1}\cdots x_{i-1}) = \sum_{j=0}^{k}\lambda_j^x P_{MLE}(x_i|x_{i-j}x_{i-j+1}\cdots x_{i-1})$$

ただし $0 \le \lambda_j^x \le 1$, $\sum_{j=0}^{k}\lambda_j^x = 1$

係数 λ の値は、削除補間法で求める。つまり、パラメータ推定のためのコーパスを k 個に分割し、$k-1$ 個の部分で確率値 P_{MLE} を推定し、残りの部分で補間係数を最尤推定するということを k 通りに渡って行い、その平均値をとる。

2.4 既知形態素と既知文字の選択方法

既知形態素や既知文字の定義は任意であるが、削除補間の考え方の延長として以下のように定義する。

> 学習コーパスを k 個に分割し、i 番目の部分コーパスをテストコーパスと見立て、残りの部分を学習コーパスと見立てた場合に、テストコーパスにのみ現れる形態素 (文字) を未知形態素 (文字) とし、それ以外を既知形態素 (文字) とする。

したがって、既知形態素と既知文字はそれぞれ以下のように定義される。

既知形態素 学習コーパスを k 個の部分コーパスに分割し、2 つ以上の部分コーパスに出現する形態素

既知文字 各部分学習コーパスの未知形態素 (重複を許す) を文字 n-gram モデル推定のための k 個の学習コーパスと見做し、これらの部分コーパスの 2 つ以上に出現する文字

3 分野適応

この節では、十分な量のコーパスが用意できない分野に対する言語モデルの予測力を、新聞記事などのような容易に入手可能な大規模コーパスを用いて向上させる方法を提案する。以下では、まず既に提案されている補間を用いる方法を説明し、次に形態素クラスタリングを用いる提案手法を説明する。以下では、形態素 2-gram モデルを用いてこれらの方法を説明するが、両手法ともにその一般形である形態素 n-gram モデルや、その他の言語モデルに容易に拡張可能である。

3.1 補間による分野適応

一般的な分野適応の手法として、対象分野の小量のコーパスから推定された言語モデルと新聞などの大規模なコーパスから推定された言語モデルの補間を線形補間する方法がある。文を形態素列とみなし、これを文頭から順に予測するモデルの場合、対象分野のコーパス (Task) から推定された言語モデルを $P_t(m_i|m_1m_2\cdots m_{i-1})$ とし、対象分野のコーパスと新聞記事などの大規模なコーパス (Base) の和から推定された言語モデルを $P_{t+b}(m_i|m_1m_2\cdots m_{i-1})$ とすると、分野適応された言語モデル $P_a(m_i|m_1m_2\cdots m_{i-1})$ は以下の式で表される (図 1 参照)。

$$P_a(m_i|m_1m_2\cdots m_{i-1}) = \lambda P_{t+b}(m_i|m_1m_2\cdots m_{i-1}) + (1-\lambda)P_t(m_i|m_1m_2\cdots m_{i-1}) \text{ ただし } 0 \le \lambda \le 1$$

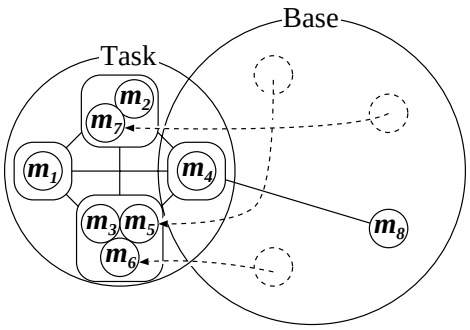


図 2: クラスタリングによる分野適応

補間係数 λ は、対象分野のコーパスの出現確率が最大となるように削除補間法で推定される。

3.2 形態素クラスタリングによる分野適応

対象分野の学習コーパスが小量しかない場合、その分野のコーパスに出現する可能性が十分にあるにもかかわらず、実例が観測されない形態素 (1-gram) や形態素列 (n-gram, $n \geq 2$) が多数存在することになる。これらのいくつかはテストコーパスに出現し、結果として言語モデルの予測精度の低下を招く。分野適応の目的は、このような形態素や形態素列の数を減少させることで予測精度を改善することである。前項で説明した補間による方法では、形態素単体での出現の予測には有効であるが、対象分野における履歴を考慮して出現を予測することはできない。一般コーパスに含まれる語彙の予測に履歴が活用されるのは、履歴が一般コーパスの語彙の列である場合のみであり、対象分野のテストコーパスにこれが現れる頻度は必然的に低くなる。

本論文で提案する分野適応の基本となる考え方は、一般コーパスの各語彙の対象分野の文脈での振る舞いの予測に、対象分野の語彙の振る舞いを利用することである。この実現方法の概略は以下の通りである (図 2 参照)。

1. 対象分野の学習コーパスから形態素 2-gram モデルを推定する。
2. これをそのままクラス 2-gram モデルとみなす (1 つの形態素に 1 つのクラスが対応する)。
3. 一般コーパスから得られる形態素を一般コーパスにおいて類似の振る舞いをする対象分野の形態素のクラスに、これが見つかる限りにおいて併合する (詳細後述)。
4. 併合するクラスが見つからなかった一般コーパスの形態素は個別のクラスとする。
5. このようにして得られたクラス分類を用いて一般コーパスのクラスの 1-gram と 2-gram を計数する。
6. 対象分野のコーパスを対象として計数したクラスの 1-gram と 2-gram と、一般コーパスを対象として計数したクラスの 1-gram と 2-gram を補間し、クラス間の遷移確率を得る。
7. 各クラスからの形態素の出現確率を計算する (詳細後述)。

以上の方法で得られる言語モデルは、クラス 2-gram モデルの一種である。したがって、各形態素の予測は、以下の式が示すように、クラス列から次のクラスを予測する部分と予測されたクラスからさらに形態素を予測する部分との積で表される。

$$P(\boldsymbol{m}) = \prod_{i=1}^{n} P(m_i|c_i)P(c_i|c_{i-1})$$

以下では、上式の 2 つの条件付き確率の計算方法を順に詳述する。

予測されたクラスからの形態素の予測は、以下の式が示すように、クラスに含まれる形態素が唯一の場合と、複数ある場合に分けられる。これが複数ある場合には、対象分野のコーパスから得られた形態素の場合と、一般コーパスから併合された形態素の場合に分けられる。ただし、対象分野のコーパスから得られる語彙を $\mathcal{M}_t$、クラスを $\mathcal{C}_t$ とし、対象分野のコーパスから得られるクラスから対象分野のコーパスから得られる語彙が出現する確率を α とする。

$$P(m_i|c_i) = \begin{cases} \alpha & \text{if } c_i \in \mathcal{C}_t, m_i \in \mathcal{M}_t \\ (1-\alpha)\frac{f_b(m_i)}{\sum_{m \in c_i} f_b(m)} & \text{if } c_i \in \mathcal{C}_t, m_i \notin \mathcal{M}_t \\ 1 & \text{otherwise} \end{cases}$$

式中の $1-\alpha$ は、クラスタリングによって併合された一般コーパスの形態素に割り当てられる出現確率であり、これが一般コーパスからの各形態素に一般コーパスでの出現頻度に比例して配分される。この α は、以下の式のように決定される。

$$\alpha = \frac{\sum_{c_i \in \mathcal{Y}} f_t(y(c_i))}{\sum_{c_i \in \mathcal{Y}} \sum_{m \in c_i} f_t(m)}$$

ここで、$\mathcal{Y} \subseteq \mathcal{C}_t$ は一般コーパスから併合された形態素が存在するクラスの集合を表し、$y(c)$ はクラス c の対象分野のコーパスから得られる形態素を表し、f_t は対象分野の学習コーパスでの頻度を表すとする[1] 。

次のクラスを予測する部分は、以下の式が示すように、対象分野のコーパスのクラス 1-gram モデルとクラス 2-gram モデルと、対象分野のコーパスと一般コーパスの和のクラス 1-gram モデルとクラス 2-gram モデルとの補間とする。

$$P(c_i|c_{i-1})$$

[1] 一般コーパスの語彙の一部は、$f_t(m) \geq 1$ となることに注意 (前節の既知語選択の方法参照)。

$$= \begin{cases} \lambda_1 P_t(c_i) + \lambda_2 P_s(c_i) + \lambda_3 P_t(c_i|c_{i-1}) + \lambda_4 P_s(c_i|c_{i-1}) \\ \qquad\qquad \text{if } f_t(c_{i-1}) > 0 \\ \lambda_5 P_t(c_i) + \lambda_6 P_s(c_i) + \lambda_7 P_s(c_i|c_{i-1}) \ \text{otherwise} \end{cases}$$

この式で、P_t は対象分野のコーパスから最尤推定された確率を表し、P_s は対象分野のコーパスと一般コーパスの和から最尤推定された確率を表す。

3.3 分野適応のための形態素クラスタリング

前節で述べたように、提案する分野適応手法の中心は、一般コーパスにのみ出現する形態素の文脈情報として、適応する分野の類似の語彙の文脈情報を利用することである。これを実現するために、適応する分野の語彙をクラスとみなし、一般コーパスにのみ出現する形態素をこれらの中から振舞いの類似という観点で最適のクラスに併合する。どのクラスが最適であるか、あるいは最適なクラスがないかは、形態素クラスタリング [5, 6, 7, 8] によって計算する。次節で述べる実験では、削除補間を応用となる平均クロスエントロピーを基準とし、解をボトムアップで探索する形態素クラスタリング [7] を用いた。つまり、一般コーパスにのみ出現する形態素を頻度の降順に並べ、この順に各形態素を適応する分野の各形態素を移動対象のクラスとして併合した場合の効果を計算する。併合した場合の効果は、適応する分野のコーパスと一般コーパスの和に対する平均クロスエントロピーの減少量で、この減少幅が最も大きいクラスに移動する。いかなる移動の場合にも平均クロスエントロピーが増加する場合は、移動しないでその形態素のみからなるクラスとする。

4 評価

3 節で述べた言語モデルの分野適応の効果を既存の線形補間やコーパスを単純に足す方法などと比較する実験を行なった。この節では、実験の条件と結果を提示し、提案手法の評価について述べる。

4.1 実験の条件

実験に用いたコーパスは、対象分野の小量のコーパスとしての放送大学の書き起こしと、大規模な一般コーパスとしての日本経済新聞の記事である。各文は、単語に分割され、各単語には品詞が付与されている。放送大学の書き起こしは 10 個に分割し、1 つをテストコーパス、9 つを学習コーパスとした。日本経済新聞の記事は 9 つに分割した。これら 9 つに分割された学習コーパスから、削除補間に基づく語彙選択やクラスタリングを行なった (表 1 参照)。以下の説明では、9 分割された放送大学の学習コーパスと日本経済新聞の学習コーパスのそれぞれ 1 つずつを足し合わせることで得られる 9 つの学習コーパスを単純和コーパスと呼ぶ。

表 1: コーパス

	文数	形態素数	文字数
学習 (日経新聞)	10,098	302,297	462,130
学習 (放送大学)	990	42,243	61,812
テスト (放送大学)	110	4,455	6,561

4.2 各モデルの詳細

実験では、提案手法の有効性の確認を目的として、以下のモデルを構築し、それぞれの予測力を比較した。

- コーパス単純和モデル (頻度の和)

  単純和コーパスから 2 節で説明した方法で形態素 2-gram モデルを構築する。

- 線形補間による分野適応 (既存手法)

  単純モデルの既知形態素を用いて、放送大学の学習コーパスから 2 節で説明した方法で構築した形態素 2-gram モデルと、単純和コーパスから同様の方法で得られる形態素 2-gram モデルを 3 節で説明した方法で線形補間する。

- クラスタリングによる分野適応 (提案手法)

  単純モデルの既知形態素を用いて、単純和コーパスから 2 節で説明した方法で構築した形態素 2-gram モデルを構築する。放送大学の学習コーパスから得られる語彙を既存クラスとし、それ以外の頻度 5 以上の既知形態素を対象として、3 節で説明した形態素クラスタリングを実行する。

各分野適応モデルの未知語モデルは単純モデルと同じとする。既知形態素は共通なので、未知語の表記の予測によるエントロピーへの寄与は各モデルで一定である。

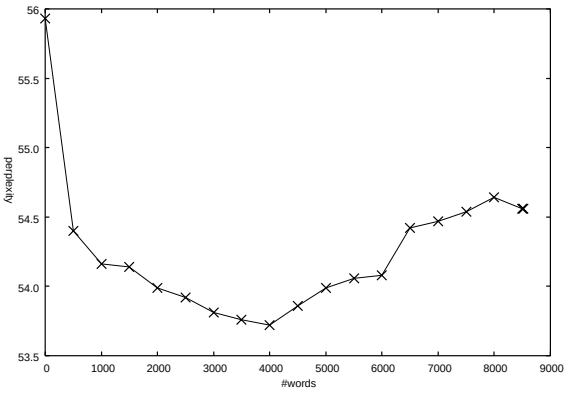
4.3 評価

各言語モデルの予測力を評価するために、対象分野のテストコーパスに対する文字あたりのエントロピーと形態素あたりのパープレキシティーを計算した。表 2 は、この結果である。パープレキシティーの比較の結果、線形補間よる分野適応の結果得られるモデルのパープレキシティーは、コーパスを単純に足すことで得られるモデルのパープレキ

表 2: 各モデルの予測精度 (文字単位のエントロピー)

適応方法	形態素の予測	未知語の表記の予測	合計	PP
頻度の和	3.865	0.511	4.376	87.11
線形補間	3.492	同上	4.003	59.52
クラス LM	3.391	同上	3.902	53.72

※PP は形態素単位のパープレキシティ
クラス LM の値は頻度 5 以上の形態素を対象にした場合



対象形態素数が 0 の場合の値 (55.93) は、両コーパスの 1-gram と 2-gram の 4 つを同時に補間した場合のパープレキシティである。

図 3: クラスタリング対象の形態素数と予測力の関係

シティーよりも大幅に低くなっており、既存の線形補間による方法で予測力が大きく改善することが確認された。クラスタリングによる分野適応の結果得られるモデルのパープレキシティーは、既存の線形補間の結果得られるモデルパープレキシティーよりも 9.74%低く、提案手法が線形補間による方法よりも優れていることが確認された。

以上の実験でのクラスタリングは、頻度 5 以上の 3,779 個の形態素に対して頻度の降順に行なった。さらに、クラスタリングの対象とする形態素の数を変化させた場合のパープレキシティーを計算し、図 3 のグラフを得た。この結果から、一般コーパスの形態素を適応分野の語彙に併合する効果は、頻度の低下にともなって減少し、4,000 語 (頻度 4) 付近で負の効果を持つようになる。この結果は、頻度が低い形態素のクラスタリングが不正確であるというクラスタリングの知見と合致する。また、クラスタリングの評価関数が、適応分野のコーパスでの期待予測力 (クロスエントロピー) ではなく、一般コーパスでの期待予測力とせざるを得ないこともこの原因であろう。

5 結論

本論文では、確率的言語モデルを大規模コーパスがない分野に適応する方法を提案し、その有効性を実験的に示した。提案手法では、適応する分野の小規模コーパスには出現しない語彙であっても、一般的な大規模コーパスに出現すれば、適応する分野から得られる文脈情報も予測に用いることを可能にする。これは、適応する分野の単語をクラスとみなし、一般的な大規模コーパスに出現する語彙を類似のクラスに割り振ることで実現される。

実験では、コーパスとして放送大学の書き起こしと日経新聞の記事を用いて、従来の線形補間による方法と提案手法の予測精度を比較した。パープレキシティーの比較の結果、提案手法によるモデルは線形補間によるモデルに対して 9.71%の改善が確認された。この結果、音声認識を代表とする確率的言語モデルの応用において、本論文で提案する分野適応の方法が、従来手法に比べて有効であることが確認された。

参考文献

[1] Shoichi Matsunaga, Tomokazu Yamada, and Kiyohiro Shikano. Task adaptation in stochastic language models for continuous speech recognition. In *Proc. of the ICASSP92*, pp. 165–168, 1992.

[2] Reinhard Kneser and Volker Steinbiss. On the dynamic adaptation of stochastic language models. In *Proc. of the ICASSP93*, pp. 586–589, 1993.

[3] P. R. Clarkson and A. J. Robinson. Language model adaptation using mixtures and an exponentially decaying cashe. In *Proc. of the ICASSP97*, pp. 799–802, 1997.

[4] Fredelick Jelinek, Robert L. Mercer, and Salim Roukos. Principles of lexical language modeling for speech recognition. In *Advances in Speech Signal Processing*, chapter 21, pp. 651–699. Dekker, 1991.

[5] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jennifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, Vol. 18, No. 4, pp. 467–479, 1992.

[6] Hermann Ney, Ute Essen, and Reinhard Kneser. On structuring probabilistic dependences in stochastic language modeling. *Computer Speech and Language*, Vol. 8, pp. 1–38, 1994.

[7] Shinsuke Mori, Masafumi Nishimura, and Nobuyasu Itoh. Word clustering for a word bi-gram model. In *ICSLP*, 1998.

[8] Jianfeng Gao, Joshua Goodman, Guihong Cao, and Hang Li. Exploring asymmetric clustering for statistical language modeling. In *Proc. of the ACL02*, pp. 183–190, 2002.