音声言語情報処理 28－3
（1999. 10. 29）

SVDを用いたトライグラム行列の分析

寺島 志郎† 武田 一哉†* 板倉 文忠‡*

†名古屋大学大学院工学研究科
‡名古屋大学情報メディア教育センター
*統合音響情報研究拠点

〒464-8603 名古屋市千種区不老町1
terasima@itakura.nuee.nagoya-u.ac.jp

あらまし 単語 N-gram モデルには, その構築にあたり広大な記憶容量が必要となる問題がある. そのため, パラメータの情報を圧縮することが要求される. その解決策の1つとして, 単語クラスタリングが提案されている. 本稿では, 山本らの接続の方向性を考慮した多重 N-gram モデルの考え方と同じように, 文脈における単語の接続性を考慮した Trigram モデルのためのクラスタリング手法を提案する. 本手法は, 特異値分解 (SVD) により先行する単語をベクトル空間上に配置し, その位置関係をクラスタリングの基準とするものである. 相互情報量最大化に基づく手法により得られたクラスと3つ組のパラメータ数で比較した場合より少ないパラメータ数で良いモデルを得ることができた.

キーワード 特異値分解 (SVD), 次元数, クラスタリング, エントロピー

The Analysis of Trigram Matrix by SVD

Shiro TERASHIMA† Kazuya TAKEDA†*
Fumitada ITAKURA‡*

†Graduate School of Engineering, Nagoya University
‡Center for Information Media Studies, Nagoya University
* Center for Integrated Acoustic Information Research

Furo-cho 1, Chikusa-ku, Nagoya 464-8603 JAPAN
terasima@itakura.nuee.nagoya-u.ac.jp

Abstract Word-based N-gram model has a problem that it requires an enormous number of parameters. Thus, with limited number of training data, it is important to reduce the parameter of the model. In this paper, we will propose a vector representation of N-gram language models through SVD analysis of tri-gram matrix. In the vector space, tri-gram matrix can be represented by less than 5through a word clustering experiment where the class-based tri-gram given in the vector space has lower perprexity than clustering results of the Maximum Mutual Information approach.

Key words Singular Value Decomposition, Rank, Clustering, Entropy

1 はじめに

現在, 単語 N-gram モデルは連続音声認識システムの言語モデルとして広く用いられている. これは, 統計的な情報を基に構文情報を与え, 言語を推定するもので, 非常に有効なモデルとされている. しかし, N-gram 言語モデルに必要なパラメータ量は語彙数の 3 乗に相当する. このため, 学習データが十分に確保できない場合や, 充分な記憶容量が利用できない場合など, 応用分野によってはパラメータ空間の圧縮が不可欠である. これは N-gram モデルが N 組単語の共起に関する単純に統計に基づいており, 言語の構造に関するモデルを陽に持たないためと考えられる.

本稿では, N 組単語の共起に関し, 線形空間モデルを仮定することで, N-gram 言語確率の効率的な表現方法を検討する. 検討する手法は, N 組単語の共起頻度行列に SVD 分析法を適用することを基本とする. この結果, 単語, 構文状態それぞれが, K 次元空間上のベクトルとして表現され, それらの間に距離を定義することが, 可能になる. さらに, 定義された距離に基づき構文状態のクラスタリングを行うことで, N-gram 確率パラメータ空間の圧縮を試みる.

2 N-gram 共起頻度行列の特異値分解

2.1 構文状態のベクトル表現

構文状態 g_i において単語 w_j が出力される頻度 $F(g_i, w_j)$ を要素とする行列 $X_{ij} = F(g_i, w_j)$ を考える. 今, 構文状態として直前 $(N-1)$ この単語連鎖を考える N-gram 言語確率においては, 語彙の大きさを M とすると, 考えうる $(N-1)$ 組単語の組合せ総数は $L = M^{N-1}$ となり, 行列 $\boldsymbol{X}$ は, L 行 M 列の行列として表現される. (例えば, 本稿の実験では, $M = 2000$ のトライグラムを考えているため, $\boldsymbol{X}$ は 400 百万 × 2000 の行列となる.) この行列の第 i 行ベクトル $[F(g_i, w_1), \cdots, F(g_i, w_M)]$ は, 構文状態 g_i において, 単語 $w_1, \cdots, w_M$ が出力される (後続する) 頻度, すなわち構文状態の言語的な性質を M 次元ベクトル空間上で表現している.

一方, 行列 $\boldsymbol{X}$ の第 j 列ベクトル $[F(g_1, w_j), \cdots, F(g_L, w_j)]^T$ は, 単語 w_j が構文状態 $g_1, \cdots, g_L$ において出力される頻度, すなわち単語の言語的な性質を, L 次元ベクトル空間上で表現している. (ただし $(\cdot)^T$ で転置を表す.)

2.2 特異値分解

共起頻度行列の特異値分解は下図の式により与えられる.

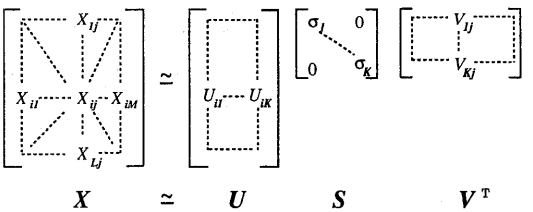

図 1: 特異値分解

ここで, K は行列 $\boldsymbol{X}$ の (打ち切り) 次元数, $(\cdot)^{\mathrm{T}}$ は転置, σ_k は特異値で特異値間には $\sigma_1 \geq \cdots \geq \sigma_K$ の関係がある. また, $\boldsymbol{u}_i$, は K 次元空間上に射影された, 先行する単語列 g_i のベクトルである. 同様に $\boldsymbol{v}_j$ は K 次元空間上に射影された, 単語 w_j のベクトルである. また, $\boldsymbol{U}$ はベクトル $\boldsymbol{u}_i$ の行列, $\boldsymbol{V}$ はベクトル $\boldsymbol{v}_i$ の行列, $\boldsymbol{S}$ は特異値の対角行列である.

SVD, 及び K 次元での打ち切り操作により, サイズ M^N の行列 $\boldsymbol{X}$ が, 行列 $\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{V}$ 各々のサイズ $(M^{N-1} \times K), K, (M \times K)$ で近似的に表現することが可能となり, 頻度行列 $\boldsymbol{X}$ の保持に必要なパラメータ数の数は

$$\frac{K}{M} = \frac{\text{打ち切り次元数}}{\text{語彙サイズ}}$$

に圧縮される.

この操作によれば, それぞれ $L(= M^N)$ 次元, M 次元空間上のベクトルとして表現された構文状態 g_i, 単語 w_j の言語的な性質を, それぞれ K 次元空間ベクトル $\boldsymbol{u}_i$, $\boldsymbol{v}_j$ よりコンパクトに表現することが可能となる. この時, 打ち切り次元数 K に対して, $\boldsymbol{u}_i$, $\boldsymbol{v}_j$ は最小 2 乗誤差 $||\boldsymbol{X} - \hat{\boldsymbol{X}}||_F^2$ を与えることが保証される.

2.3 クラスタリング

K 次元空間上に配置された構文状態 g_i のベクトル $\boldsymbol{u}_i$ から, 空間上の近さを基準にクラスタリングを行なう. その際の類似度の尺度には, *cosine* 関数

$$cosine(\boldsymbol{u}_i, \boldsymbol{S}, \boldsymbol{u}_j) = \frac{\boldsymbol{u}_i \boldsymbol{S}^2 \boldsymbol{u}_j{}^{\mathrm{T}}}{||\boldsymbol{u}_i \boldsymbol{S}||||\boldsymbol{u}_j \boldsymbol{S}||}$$

を用いた. この距離を基に k-means 法によりクラスタリングを行なう. 構文状態のクラスを用いたクラス Trigram 確率は, 以下により計算する.

$$P(w_j|g_i) \Rightarrow P(w_j|C(g_i))$$

表 1: 学習テキスト

内容	毎日新聞 93 年 1 年分
総文章数	648,205
総単語出現数	17,443,322
語彙サイズ	122,883

表 2: 評価テキスト

内容	毎日新聞 93 年 1 年分
総文章数	12,365
総単語出現数	132,933
語彙サイズ	1,983

ここで, $C(g_i)$ は構文状態 g_i が属するクラスである.

3 実験

学習テキストから Trigram 行列を生成し, 特異値分解を行なった. その結果から得られる特異値, 特異ベクトルを基に 2 種類の実験を行なった.

1 つは, 次元数と特異値と特異値ベクトルから求められる近似行列 $\hat{\boldsymbol{X}}(=\boldsymbol{USV}^T)$ の精度の関係の調査で, 近似行列と元の行列との誤差及び固有値の累積寄与率を評価した. また, 近似行列の値を基に単語 Trigram を構築し, そのモデルのエントロピーと次元数の関係も調査した.

もう 1 つは, 構文状態のクラスタリングで, 特異値と特異ベクトルから, K 次元空間上でのベクトルを基に前節で述べたコサイン関数を類似度尺度にクラスタリングを行ない, 次元数とクラスタリングの性能をクラスモデルのエントロピーにより評価した. また, 予備実験及び比較のために Bigram でも同様の実験を行なった.

3.1 実験条件

Trigram 行列の生成及びクラス Trigram モデルの学習には表 1の内容のテキストを用いた. その際, 出現頻度上位 2,000 を語彙とし, その他を未知語とした. 未知語は語彙として扱わなかった. 語彙サイズ 2,000 は小さいが, 総単語出現数に占める 2,000 語彙の単語の割合は 78(%) であり結果は有用であると考えられる. また, 評価テキストには, 語彙とした単語のみで構成されている文章を学習テキストから抜きだして用いた. よって, 評価テキストはクローズドなものである.

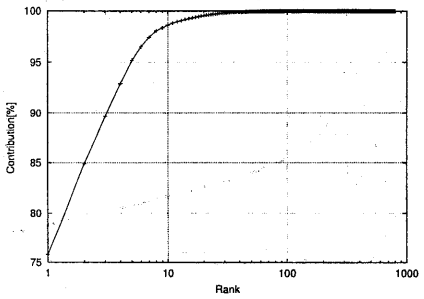

図 2: 累積寄与率 (Bigram)

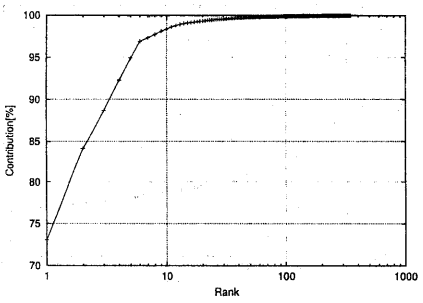

図 3: 累積寄与率 (Trigram)

3.2 Trigram 行列の特異値分解

表 1の学習テキストから Trigram 行列を生成した. その際, 2,000 語彙で構成されている 3 つ組のみを考慮した. 先行する単語列の種類の総数は 225,502 になり, Trigram 行列は 225,502 × 2,000 の行列である. また, 非零の要素の数は 1,469,722 で行列の密度は 0.33(%) であった. Bigram 行列は 2,000 × 2,000 の行列で行列の密度は 5.64(%) であった.

それぞれの行列をブロック Lanczos 法により Trigram は 350 次元, Bigram は 800 次元まで特異値分解を行なった.

3.3 累積寄与率

特異値より次元数と累積寄与率

$$\frac{(\sigma_1^2+\sigma_2^2+\cdots+\sigma_K^2)}{||\boldsymbol{X}||_F^2}$$

の関係を調査した. ここで, $||\boldsymbol{X}||_F$ は行列 $\boldsymbol{X}$ のフロベニウスノルムである. Bigram における結果を図 2に Trigram における結果を図 3に示す. 累積寄与率の値は次元数が大きくなるにつれて, 最初は急激に, 後は緩やかに上昇する. Trigram では, 10 次元で 98.39(%), 100 次元で 99.90(%), 350 次元で 99.98(%) になった. Bigram

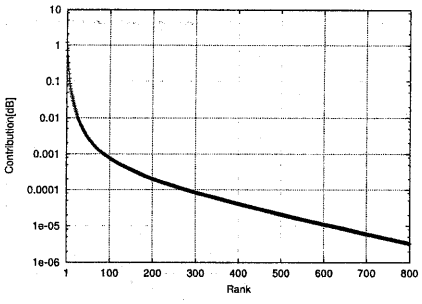

図 4: 累積寄与率-dB(Bigram)

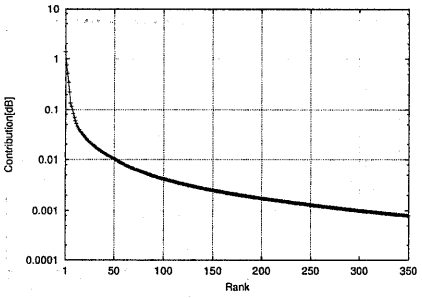

図 5: 累積寄与率-dB(Trigram)

では, 100 次元で 99.97(%), 800 次元で 99.99(%) であった. また,

$$-10\log_{10}\left(\frac{\sigma_1^2+\sigma_2^2+\cdots+\sigma_K^2}{||\boldsymbol{X}||_F^2}\right) \qquad (1)$$

として累積寄与率を調べた. Bigram の結果を図 4に, Trigram の結果を図 5に示す. どちらの場合も次元数が大きくなるに連れて 0 に漸近していく. しかし, Bigram と比較して Trigram における値はまだまだ大きく, 収束する速度がかなり遅いことが分かる.

3.4 誤差

元の行列 $\boldsymbol{X}$ と近似行列 $\hat{\boldsymbol{X}}$ との誤差を式 (2) により定義し, 次元数との関係を調べた.

$$-\frac{1}{N}\sum_{i,j}||\boldsymbol{X}_{ij}-\hat{\boldsymbol{X}}_{ij}||^2=-\frac{1}{N}||\boldsymbol{X}-\hat{\boldsymbol{X}}||_F^2 \qquad (2)$$

ここで $N = 225,502 \times 2,000$ である.

Bigram における結果を図 6に, Trigram における結果を図 7に示す. Bigram において誤差は次元数の増加とともに 0 に漸近していく. Trigram においても 0 に漸近していくが, 速度が遅く収束するにはかなりの次元数が必要だと考えられる. また, 誤差と累積寄与率が対の関係にあることが図より分かる.

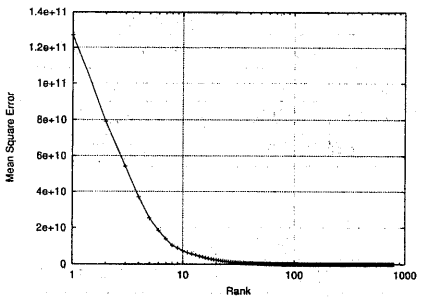

図 6: 2 乗誤差 (Bigram)

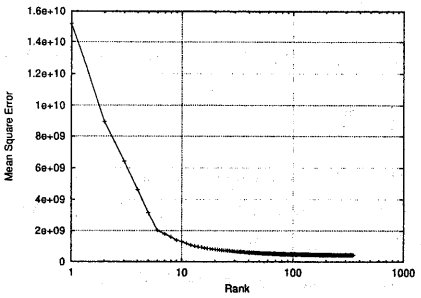

図 7: 2 乗誤差 (Trigram)

3.5 次元数とエントロピー

近似行列 $\hat{X}$ の値を Trigram の出現頻度 $\hat{F}(g_i, w_j)$ と考え, 単語 Trigram モデルを構築した. このモデルを表 2の評価テキストセットより 1 単語あたりのエントロピーにより評価し, 次元数との関係を調べた.

Trigram 確率値 $P(w_j|g_i)$ は近似行列から得られる $\hat{F}(g_i, w_j)$ を基に行ごとの総和から $\hat{F}(g_i)$ を求め, $P(w_j|g_i) = \hat{F}(g_i, w_j)/\hat{F}(g_i)$ として計算した. その際, 本来値があるのに 0 になってしまった場合, すなわち $\boldsymbol{X}_{ij} > 0$, $\hat{\boldsymbol{X}}_{ij} \leq 0$ の場合, $\hat{\boldsymbol{X}}_{ij} = 1$ とした. 確率値の平滑化は行なっていない.

Bigram の結果を図 8に, Trigram の結果を図 9に示す. どちらの場合も次元数が増加するにしたがって急激にエントロピーは減少する. Bigram において元の行列を用いた場合のエントロピーは 5.13 で, 近似行列によるエントロピーの値は, 800 次元で 5.17 となった. Trigram では, 元の行列を用いた場合のエントロピーは 3.82 で, 近似行列によるエントロピーの値は, 10 次元で 6.97, 100 次元で 5.02, 350 次元で 4.42 となった. しかし, 次元数が大きくなるにつれてエントロピーの減少は小さくなるので, 元の値に収束するにはかなりの次元数が必要と考えられる.

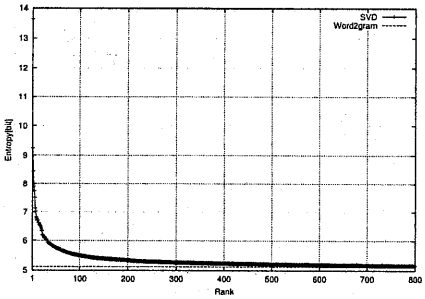

図 8: エントロピー (Bigram)

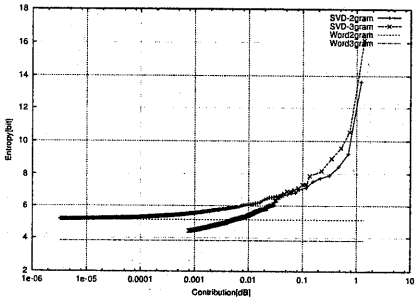

図 10: 累積寄与率-dB とエントロピー

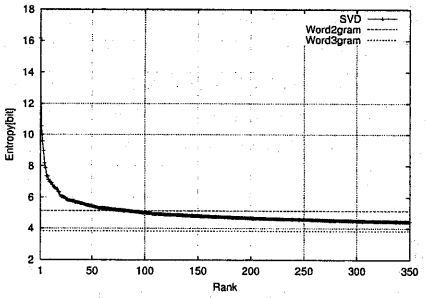

図 9: エントロピー (Trigram)

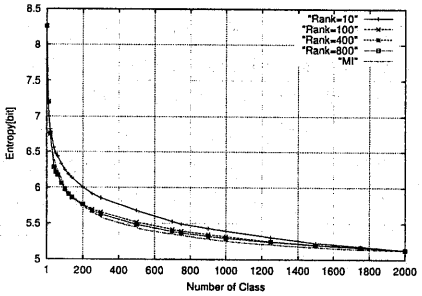

図 11: クラス数とエントロピーの関係 (Bigram)

また, 前節で述べた累積寄与率のデシベル表現とエントロピーの関係を調べた.

結果を図 10に示す. 累積寄与率のデシベル表現の値が小さくなるにつれてエントロピーの値は小さくなるが、次元数, 累積寄与率の関係を考えると, 次元数が大きくなるにつれて累積寄与率のデシベル表現の減少幅も小さくなる. このことからも Trigram では, 相当大きな次元数にならないとエントロピーの値は元の行列の値に収束しないと考えられる.

3.6 クラスタリング

3.6.1 クラス数とエントロピーの関係

Bigram モデルにおいて, クラスタリングを行ないクラス数とエントロピーの関係を調査した. 次元数は 10, 100, 400, 800 について行なった. Bigram においては, 先行する単語と後続する単語をそれぞれ 2.3 節で述べた手法によりクラスタリング, そのクラス間の遷移を考慮したクラスモデルを生成し, モデルの性能を表 2を用いてエントロピーにより評価した. クラス Bigram 確率は単語 Bigram 確率を

$$P(w_j|w_i) \Rightarrow P(w_j|C_w(w_j))P(C_w(w_j)|C_g(w_i))$$

として計算する. その際, $C_g(w_i)$ は先行する単語 w_i が属するクラスであり, $C_w(w_j)$ は後続する単語 w_j が属するクラスである.

また, 相互情報量最大化に基づく手法 [4] により得られたクラスを基にしたクラス Bigram モデルとの比較を行なった. 相互情報量最大化に基づく手法は, Bigram の共起関係を基にした手法で, 単語を接続性を考慮することなく, 1 つのクラスにクラスタリングするものである. この時, クラス Bigram 確率は

$$P(w_j|w_i) \Rightarrow P(w_j|C_{MI}(w_j))P(C_{MI}(w_j)|C_{MI}(w_i))$$

として計算する. ここで $C_{MI}(w_i)$ は単語 w_i が属するクラスである.

結果を図 11に示す. 次元数間の違いを見てみると, 次元数 10 の場合にはエントロピーの値は大きくなったが, それ以上の次元数では, ほぼ同じだった. また, 相互情報量最大化に基づく手法と同じクラス数で比較した場合, 若干エントロピーの値は大きくなった.

Trigram においても 2.3 節で述べた手法により, 先行する単語列を次元数 10, 100, 350 の場合についてクラスタリングを行った. 求めるクラス数は, 100, 500, 1000, 5000, 10000, 50000, 100000 とした. 得られたクラスを基にクラスモデルを生成し, 表 2を用いてエントロピー

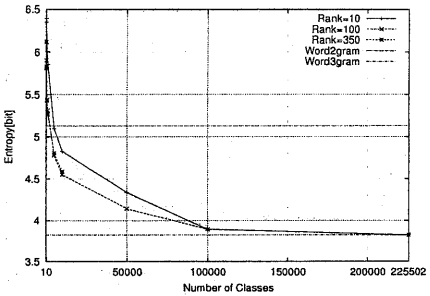

図 12: クラス数とエントロピーの関係 (Trigram)

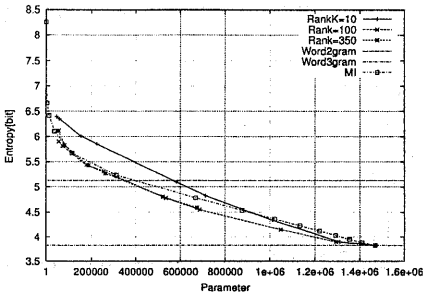

図 13: パラメータ数とエントロピーの関係

により評価した.

結果を図 12に示す. 次元数とクラスタリングの関係を見ると, 次元数が 10 と他の 2 つとの間に差はあったが, 次元数 100, 350 の間に差は見られなかった.

これから, クラスタリングにはあまりたくさんの次元数は必要ないことが分かった. 次元数 100, 350 では構文状態クラスのクラス数が 2000, 次元数 10 ではクラス数が 6000 を越えると単語 Bigram モデルよりも性能の良いモデルが得られた.

3.6.2 パラメータ数とエントロピーの関係

記憶すべき 3 つ組のパラメータ数とエントロピーに関係について調査した. その際, 相互情報量最大化に基づく手法により得られたクラスを基にしたクラス Trigram モデルとの比較を行なった. このクラスを用いたクラス Trigram モデルはクラス間の遷移を考慮するもので, クラス Trigam 確率は, 単語 Trigram 確率を

$$P(w_j|w_{j-2}, w_{j-1})$$
$$\Rightarrow P(w_j|C_{MI}(w_j))P(C_{MI}(w_j)|C_{MI}(w_{j-2}), C_{MI}(w_{j-1}))$$

として計算する. ここで, $C_{MI}(w_j)$ は単語 w_j が属するクラスである.

結果を図 13に示す. 同じパラメータ数で比較した場合, パラメータ数が多い時には相互情報量最大化に基づく手法よりエントロピーの値は小さくなり, パラメータ数が小さい時はほぼ同じだった.

4 まとめ

Trigram 行列を特異値分解し, 次元圧縮, パラメータ削減を行う実験を行なった. 特異値, 特異ベクトルから復元された近似行列において, 次元数とその精度には密接な関係があった. また, Bigram の場合と比較して, よい近似をするには大きな次元数が必要でなことが分かった. しかし, クラスタリング後の性能では, 100 次元と 350 次元とには大きな差は見られなかった. 先行する単語列のクラスタリングにおいて, パラメータ数とエントロピーの関係から相互情報量最大化に基づく手法より性能の良いクラス Trigram モデルを得ることができた. 今後は, オープンな評価テキストでクラスタリングの性能を評価したい.

謝辞

本研究の一部は文部省科学研究費補助金 COE 形成基礎研究費（課題番号 11CE2005）及び基盤 C（課題番号 11680386）の補助を受けて行われた.

参考文献

[1] S. Deerwester, S T.Dumais, R.Harshman : "Indexing by Latent Semantic Analysis," Jounal of the Society for Infomation Science 41(6), pp 633-636, 1990

[2] J. R. Bellegarda, J. W. Butzberger, Y.Chow, N. B. Coccaro, and D. Naik : "A novel word clustering algorithm based on latent semantic analysis,"In proceedings of ICASSP-96, vol.1, pp.172-175, Atlanta, May 1996.

[3] 山本, 匂坂 : "接続の方向性を考慮した多重クラス N-gram モデル," 日本音響学会平成 10 年度秋季研究発表会講演論文集分冊 I , No.2-1-19, pp.75-77

[4] P.F.Brown, V.J.D.Pietra, P.V.deSouza, J.C.Lai, R.L.Mercer : " Class-based n-gram models of natural language" ,Computational Linguistics 18, no.4, pp.467-479,1992.