自然言語処理 131－1
音声言語情報処理 26－1
（1999．5．28）

クラス統計と単語統計の併用による小規模学習データのための統計的言語モデル構成法

小林 紀彦　小林 哲則

早稲田大学 理工学部

あらまし:

エントロピーを基準として自動的にクラスタリングされたクラス情報を用いたクラス bigram と単語 bigram を、先行単語の unigram 頻度の関数として表される重みを用いて、融合することによって、少量の学習データからでも精度と信頼性を保つことのできる頑健な言語モデルを構築することを試みた。新聞記事読み上げ文の認識実験の結果、学習データ量 40000 文での融合 bigram（認識率 82.01%）が、学習データ量 100000 文での単語 bigram（認識率 82.11%）に相当することがわかった。また、学習データが少量しかないスポーツニュースでの会話文を学習データとした場合、単語 bigram と比較してクラス bigram で最大 4.2 ポイント、融合 bigram で最大 5.5 ポイントの認識率が向上した。

Roubust language modeling for small training dataset using class combined word bigram

Norihiko KOBAYASHI, Tetsunori KOBAYASHI

School of Science and Engineering, Waseda University

Abstract:

We propose a method of robust language modeling for a small amount of training data. In this method, the word bigram and the class bigram are combined using a weighting function of preceeding word freqency. We made experiments on speech recognition using Mainichi newspaper articles and JNAS speech corpus. As the results, the accuracy of class combined bigram for 40000 training text reached 82.01%, which is equivalent to that of word bigram accuracy for 100000 trainig text. We also made experiments using a small amount of training data consist of sports news dialogue on TV. Recognition accuracy of word bigram was 77.8%, class bigram was 82.1%, class combined bigram was 83.3%. The class combined bigram improved 5.5 points of recognition accuracy as compared with word bigram.

連絡先：〒 169 東京都新宿区大久保 3-4-1 早稲田大学理工学部 電気電子情報工学科　小林 哲則
TEL: (03)5286-3379 FAX: (03)3205-9381 E-mail: koba@tk.elec.waseda.ac.jp

1 はじめに

大語彙連続音声認識において、言語モデルとして単語N-gramが広く用いられている。単語N-gramは作成の簡便さや表現力の高さ（単語列に応じたきめ細かい出力確率の表現）等、優れた点が多い一方、大量の学習データを必要とするという問題点がある。このため会話音声認識のような大量の学習データの収集が難しい分野では、しばしばクラスN-gramが用いられている。一般にクラスの種類数は単語の種類数より非常に小さいため、クラスN-gramモデルを使用することで計算量や記憶領域を削減する効果がある。また、単語N-gramモデルの学習よりも小規模な学習データで、安定にN-gram頻度を推定できるという効果もある。しかし、クラスの拘束を受けるため、表現力の面では単語N-gramに比べ劣っている。

このような観点から、単語N-gramとクラスN-gram双方の利点を引き出すための統合の枠組が検討されている。代表的な方法としては、単語並びの種類に応じて、単語統計を使うのか、クラス統計を使うのかを切替える方法や[5][6]、単純な重みづけ和によって単語統計とクラス統計とを平均化する方法[7]などが試みられている。

本研究では、単語統計とクラス統計との重みつき和で統合する手法を採用するが、結合の重みを単語統計の信頼度に応じて決めることが従来の手法とは異なる。単語bi-gramの信頼度は、先行単語の頻度に応じて決まり、その値が大きいほど信頼度は高くなるものと考えられる。そこで、先行単語の頻度が大きければ、単語統計の重みが高くなるよう重み関数に拘束条件を入れた上で、この関数形を決めるパラメータを学習データから最尤推定によって求めることを試みる。また、求められた重み関数によって統合された言語モデルを用いることの効果を新聞記事の読み上げデータと、スポーツを題材にした会話音声データの認識実験により検証する。

2 bigram言語モデルの学習

連続音声認識に必要なbigram確率は、大規模なテキストデータベース（学習データ）を用いて推定されることが多い。しかし、統計的言語モデルが抱える本質的な問題として、次のようなものが挙げられる。

- 学習データに現れなかった単語列は、出現確率が0と推定され、そのような単語列が実際に発話されても認識できない。
- 学習データ（あるいは認識対象）の言語的偏向により、ある単語列の出現確率が不当に低く、または高く推定されてしまうことがある。

学習データが無限にあるなら（対象とする言語の言語的特徴をすべて含み、すべての単語連鎖について出現確率を推定するのに十分な標本数を備えているなら）、上述の問題点はクリアされ単語bigramに対するクラスbigramの優位性は見当たらなくなる。

しかし、実際には学習データは有限であり、上記の前提は成り立たない。このためクラスbigramは単語bigramに対して次のような優位性があると考えられる。

1. 単語bigramよりクラスbigramの方が組合せの数が少ないために単語bigram確率を推定する時よりもより多くの標本が得られ、推定されたbigram確率の信頼性が高い。
2. いくつかの単語をひとまとめにした単語クラスの持つ言語的特徴を扱うため、話題や時期などによる言語的偏向に対して頑健である。

特に前者は、学習データが大量に無い場合により強調される特徴であろうと考えられる。

以上の仮定を基に、本研究では単語bigramとクラスbigramを融合することによってこれらの問題を解決することを試みる。

2.1 クラスの自動分割

クラス言語モデルにおいて、クラスをどのように設定するかはモデルの性能に直結する重要な問題である。クラスの分割には、単語の品詞等の先見的知識を用いる手法と、自動的なクラスタリングに基づく手法とに分けられる。前者は、学習用のテキストが大量にない場合や認識対象文が異なるタスクの場合に比較的頑健なモデルを作成できると考えられるが、実際にはクラス分類に用いる

品詞情報は単語の性質全体を表すため、N-gramにおいて重要な単語の接続性を表しているとは言い難い。このため、前述のようにあまり良い結果を生み出すことができていない。一方後者は、クラス分類の基準を学習データ自身から作り出すために、タスクに特化した高精度なモデルが作成できるのではないかと考えられる。

本研究では、学習データとしてスポーツニュースでの会話文を用い、単語をアニーリング [10][11] に基づいて自動的に分類する手法をとった。以下に具体的なクラス分類のアルゴリズムを紹介する。

1. パラメータ cp の初期値の設定
   - 初期状態として、全単語を一つのクラスに入れておく
   - ある一つの単語に着目し、ランダムに選ばれたクラスに移動させる（設定クラス数は始めから決めておく）
   - 全ての単語も同様にクラスを移動する
   - $exp(\frac{-\text{パープレキシティの変化量の平均値}}{cp}) \simeq 1$ となるように（$= 1$ だと cp $= 0$ になってしまうので、$= 0.99$ 程度にする）cp（control parameter）の初期値を設定する
2. クラス移動の採用と棄却
   - ある単語をランダムに選ばれたクラスに移動させる
   - $exp(\frac{-\text{パープレキシティの変化量}}{cp})$ が 0 ～ 1 の任意のしきい値（この値が 1 に近いほどクラス移動の採用の条件が厳しくなり、0 に近いほど甘くなる）よりも大きければ、このクラス移動を採用、小さければ棄却する。もしクラス移動が棄却されたら、もとのクラスに戻す。
3. cp の更新 x
   - 上記の手順 2 を全単語について行ない、$exp(\frac{-\text{パープレキシティの変化量の平均値}}{cp}) \simeq 1$ から cp を更新する
4. クラス分類の継続
   - 手順 2,3 を繰り返す
5. クラス分類の終了
   - cp が一定以上小さくなったら、クラス移動の収束とみなし、クラス分類の終了とする

なお、上記のパープレキシティとは、、学習テキストにおけるパープレキシティ（トレーニングセットパープレキシティ）のことである。

本研究で採用したスポーツニュース会話文は約 4000 語彙からなっているが、分割クラス数の大小に関係なくおよそ 20 セット（全単語について 20 回ずつクラス移動の機会がある）程度で、クラス分類は収束した。また、分類終了後の分類結果を見ても、品詞や意味上の共通点を持つ単語が同一のクラスに分類されるなどといった傾向は見てとれなかった。

2.2 単語 bigram モデル

単語 bigram 確率は以下の式を用いて推定する。

$$P_w(w_2|w_1) = \frac{c(w_1, w_2)}{c(w_1)} \tag{1}$$

ここで $c(W)$ は引数で指定された単語、または単語列が学習データ中に出現した回数を表すものとする。

2.3 クラス bigram モデル

クラス bigram そのものは、単語 bigram と同様、以下のようにして推定する。

$$P(c_2|c_1) = \frac{c(c_1, c_2)}{c(c_1)} \tag{2}$$

ここで c_k は単語 w_k が属している単語クラス、$c(W)$ は引数で指定された単語または単語列が学習データ中に出現した回数である。

しかしこのままでは単語 bigram と融合する際に不都合であるので、以下のようにして単語 bigram の形式（単語から単語への遷移確率の形式）に拡張し、以後これをクラス bigram と呼ぶことにする。

$$P_c(w_2|w_1) = P(w_2|c_2)P(c_2|c_1)P(c_1|w_1) \tag{3}$$

ここで $P(c_k|w_k)$ は単語 w_k がクラス c_k に属する単語として出現する確率

$$P(c_k|w_k) = \frac{c(w_k, c_k)}{c(w_k)} \tag{4}$$

であり、$P(w_k|c_k)$ はクラス c_k を仮定した時、単語 w_k を出力する確率

$$P(w_k|c_k) = \frac{c(w_k, c_k)}{c(c_k)} \quad (5)$$

である（ただし、$c(w_k, c_k)$ は単語 w_k がクラス c_k として出現した回数を表す）。

3 単語 bigram とクラス bigram の融合

3.1 単語 bigram とクラス bigram の比較

クラス bigram は単語 bigram を縮退させたものと考えることができるが、それぞれの性質の違いとして予想されることを表 1 にまとめてみる。

表 1: 単語 bigram とクラス bigram の性質

	単語 bigram	クラス bigram
認識単位種類数	多い	少ない
表現力	高い	低い
話題への頑健性	低い	高い
確率推定精度	低い	高い

単語 bigram とクラス bigram を比較した場合、単語 bigram の方が表現力が高いと考えられるが、学習データが大量にない場合はクラス bigram の方が高精度な言語モデルとなることも多い。しかし、学習データの量が不十分なためクラス bigram を用いる場合でも、すべての単語組についてデータが不十分であるわけではなく、出現回数が多い単語については、ある程度の信頼のおける確率推定が得られると考えられる。したがって、基本的には単語 bigram を優先し、サンプル数が少なくて推定精度が落ちてしまうような単語列についてはクラス bigram を優先することを考える。

3.2 重み関数の設定

そこで言語モデルを、単語 bigram $P_w(w_2|w_1)$ とクラス bigram $P_c(w_2|w_1)$ の重みづけ和によって表し、単語 bigram にかかる重みを単語 unigram の頻度 $c(w_1)$ の関数としておき、$c(w_1)$ が大きくなるほど、重くなるように与える。この条件を満足する関数として、ここでは次式を用いる。

$$\mu(w_1) = k\left(1 - exp(\frac{-c(w_1)}{T})\right) \quad (6)$$

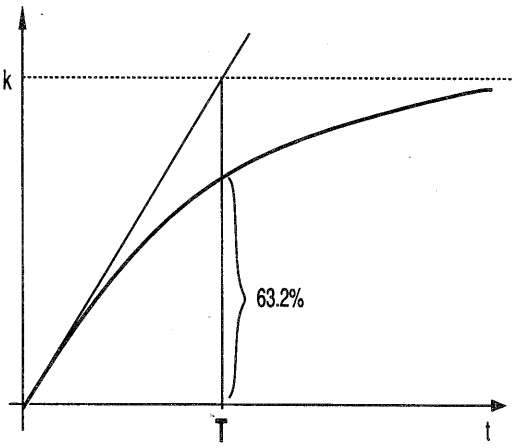

図 1: 重み関数の例

この式の利点は、パラメータ k によって収束値が、T によって応答の変化の速さ（T が大きければ、最終値に向かってゆっくり漸近し、逆に T が小さければ、早く漸近する）が変化できることである。

そしてこの重み関数を用いて融合 bigram を計算する（式 7）。

$$\begin{aligned} P(w_2|w_1) &= \mu(w_1)P_w(w_2|w_1) \\ &+ (1-\mu(w_1))\,P_c(w_2|w_1) \end{aligned} \quad (7)$$

また、本研究では、パラメータ k と T は各言語モデルセットに対して最尤推定により求めた。具体的な手法は、単語 bigram 確率 $P_w(w_2|w_1)$ とクラス bigram 確率 $P_c(w_2|w_1)$ を重みづけで融合して求まる前述の式 (7) を、別に用意したテストセットについて求め、テストセットエントロピーを計算する。

$$\sum_{w_i}^{N} logP(w_{i+1}|w_i)$$

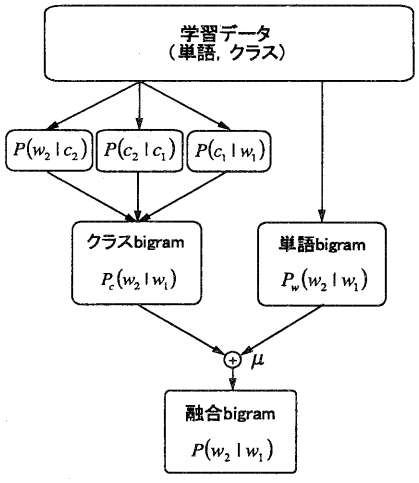

図 2: 単語 bigram とクラス bigram の融合

$$= \sum_{w_i}^{N} log\{\mu(w_i)P_w(w_{i+1}|w_i) + (1-\mu(w_i))\,P_c(w_{i+1}|w_i)\} \quad (8)$$

そして上式 (8) をパラメータ k と T で偏微分する。

$$\begin{aligned}
&\frac{\partial}{\partial k}\sum_{w_i}^{N} logP(w_{i+1}|w_i) \\
&= \frac{\partial}{\partial k}\sum_{w_i}^{N} log\left[k\left(1-exp(\frac{-c(w_i)}{T})\right)P_w \right. \\
&\quad \left. + \left\{1-k\left(1-exp(\frac{-c(w_i)}{T})\right)\right\}P_c\right] \\
&= \sum_{w_i}^{N} \frac{(P_w-P_c)\left(1-exp(\frac{-c(w_i)}{T})\right)}{\mu(w_i)P_w+(1-\mu(w_i))P_c} \\
&= 0
\end{aligned} \quad (9)$$

$$\begin{aligned}
&\frac{\partial}{\partial T}\sum_{w_i}^{N} logP(w_{i+1}|w_i) \\
&= \frac{\partial}{\partial T}\sum_{w_i}^{N} log\left[k\left(1-exp(\frac{-c(w_i)}{T})\right)P_w \right. \\
&\quad \left. + \left\{1-k\left(1-exp(\frac{-c(w_i)}{T})\right)\right\}\right] \\
&= \sum_{w_i}^{N} \frac{k\frac{c(w_i)}{T^2}exp\left(\frac{-c(w_i)}{T}\right)(P_c-P_w)}{\mu(w_i)P_w+(1-\mu(w_i))P_c} \\
&= 0
\end{aligned} \quad (10)$$

(ただし、$P_w \equiv P_w(w_{i+1}|w_i), P_c \equiv P_c(w_{i+1}|w_i)$)

以上のように偏微分した結果を 0 とすることによって、それぞれのクラス数に最適となるであろうパラメータ k と T が求まる。

4 予備実験

学習データが大量にはない場合の頑健な言語モデル構築法として、前章で説明した融合 bigram を提案したが、実際にどの程度の学習データ量の範囲で効果があるのかが知りたいところである。そこで、大量にある毎日新聞記事データを使うことによって、これらを検証した。

以下に実験諸条件を示す。

4.1 実験条件

言語モデルは、学習データとして毎日新聞記事 94 年版約 57 万文 (1 月〜9 月分) から 5000、10000、200000、30000、40000、50000 文 (単語 bigram のみ 100000 文) を任意に抽出し、それぞれから単語 bigram、クラス bigram、融合 bigram (クラス数は 500) を作成した (語彙は上位頻度 5000 語)。

音響モデルは、各音素モデルも 5 状態 3 ループの HMM で、子音は出渡りの biphone、母音は triphone になるような音素環境依存の HMM (カテゴリー数 5315) とし、特徴量は mfcc12 次元+ log パワーおよびその Δ の計 26 次元とした。

評価データは、JNAS 日本音響学会新聞記事読み上げ音声コーパスから男性話者 10 人× 10 文計 100 文 (未知語なし) を使用した。

認識アルゴリズムとしては、ワンパスビタビアルゴリズムを使用し、言語モデルと音響モデルの重みは 8:1 とした。

4.2 結果

認識結果の単語認識率を図 3 に示す。ただし、正解精度は

$$Accuracy = \frac{H-I}{N} \times 100[\%] \quad (11)$$

(H : 正解単語数, I : 挿入誤り単語数, N : 総単語数)

である。以下、「認識率」とはこの正解精度を指すものとする。

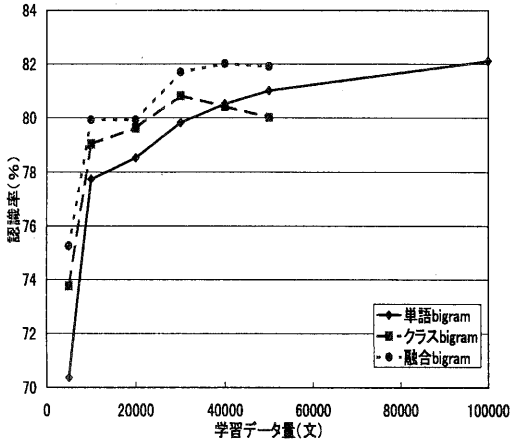

図 3: 認識結果

図3からわかるように、当然、単語bigramは学習データの増加とともに認識率の向上が伺われる。これに対してクラスbigramは学習データ30000文がピーク（認識率80.82%）となり、融合bigramでは学習データ40000文がピーク（認識率82.01%）となった。この値は、学習データ100000文の単語bigram（認識率82.11%）に匹敵する。

この結果から、学習データが多く集まらない分野に対する融合bigramの有効性が示せたといえるのではないだろうか。

5 連続音声認識実験

続いて、学習データが少量しかない場合の認識実験を行なった。

5.1 学習データ

bigram言語モデルの学習データとして、1998年8月〜11月までのTVスポーツニュース番組（NHK:サンデースポーツ、フジテレビ:プロ野球ニュース、テレビ朝日:ニュースステーションなど）をビデオに撮り、会話として成立している部分だけを人手で書き起こしたものを使用した。以下に会話文の例を示す。

表 2: 会話文の例

A: 先週は何といってもマグワイア選手の62号ホームラン。
B: そうですね、偉大な記録。
A: アメリカ中が揺れましたね。
B: そう、そして今週はもうひとり、ソーサ選手が俺を忘れてはならんと。
A: うん、終盤を盛り上げていますね。

さらにこのテキストを日本語形態素解析ツール「JUMAN version 3.5」を使用して形態素解析を行なった。解析結果はテキストが会話文ということもあり、多くの解析誤りが見られ（特に「えー」「うーん」などの会話音声特有の言い回し）、人手で修正を行なった。この結果、691会話、4471文、延べ単語数61250、語彙4043語（文頭記号 <silB>、文末記号 <silE>、ショートポーズ記号 <sp> を含む）が得られた。

この中から666会話を言語モデル作成用の学習データとして使用した。

5.2 実験条件

認識実験における諸条件は前述の通りである。言語モデルは、単語bigramとクラスbigram、融合bigram（クラス数200〜500）を作成し、単語bigramとクラスbigramを融合する際のパラメータ k と T は、それぞれの設定クラス数に対して最尤推定により求めた。クラス数250の融合bigramでは、k=0.7,T=12、となり、この時の単語bigramにかかる重みと単語種類数の関係を図4に示す。

評価データとして、学習データとは重複しない25会話（延べ単語数660語）を使用した。

5.3 実験結果

図5に認識結果を示す。

全ての設定クラスにおいて単語bigramと比較して、クラスbigramが認識率で上回り、最良で82.12%（設定クラス数250と400）と約4.2ポイ

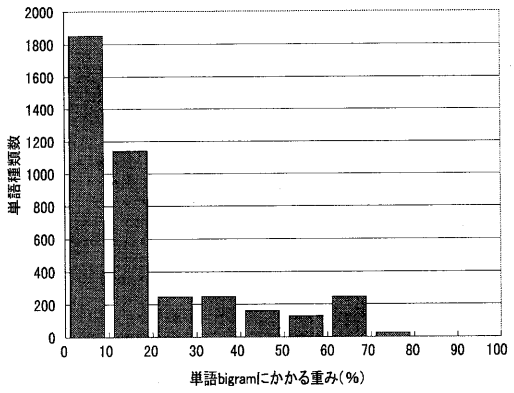

図 4: 単語 bigram にかかる重みと単語種類数の関係（クラス数 250,k=0.7,T=12 の場合）

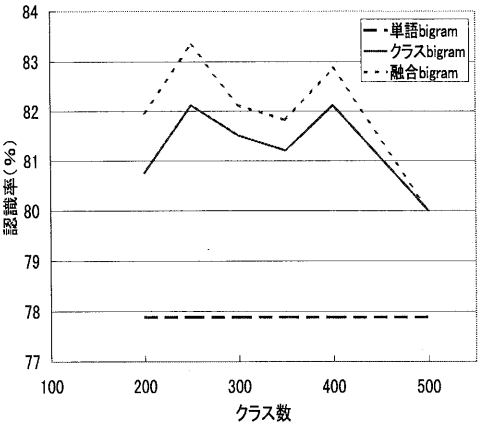

図 5: スポーツニュース会話タスクの認識結果

ントの認識率の向上となった。さらに、融合 bigram では最良で 83.33%（クラス数 250）となり、単語 bigram と比較して約 5.5 ポイントの認識率向上が得られた。これは、約 25.8%の誤認識率の改善に相当することからも、少量の学習データからの言語モデル構築法として、融合 bigram が有効な一つの手法であることがわかる。

6 まとめ

言語モデルとしてより精密である単語 bigram と、あまり精密ではないが少量の学習データからでも比較的信頼性の高いクラス bigram を融合することによって、会話タスクなど学習データが多く集まらない分野への言語モデル構成法を提案した。

タスクに毎日新聞記事を選択し、学習データ量を変化させて認識実験を行なったところ、学習データ量 40000 文から構築された融合 bigram（認識率 82.01%）が、100000 文から構築された単語 bigram（82.11%）と認識率で匹敵することがわかり、学習データが大量にはない場合の融合 bigram の有効性を示すことができた。

また、テレビのスポーツニュースでの会話文とした場合、単語 bigram と比較して融合 bigram で最大 5.5 ポイント（クラス数 250 の場合）認識率が向上した。

本研究で提案した単語 bigram とクラス bigram を融合する手法は、学習データ量が少ないほど効果が大きいことが確認されたが、学習データ量を増やしても、単語 bigram とクラス bigram それぞれの良い特徴を継承できていることもわかった。

今後は、クラスの自動分類アルゴリズムの改善や、融合の際の重み関数の設定方法など、より効果的な融合 bigram の構築に関する諸問題の検討が必要である。

謝辞

本研究には、CD-ROM 版毎日新聞記事データ、RWC テキストデータベース、JNAS 新聞記事読み上げ音声コーパスを使用した。

参考文献

[1] 肥田木康明, 小林哲則 “単語 bigram とクラス bigram の融合による話題変化に頑健な統計的言語モデル” 情処研報,vol98,No.12,pp.25-32 Feb.199 8

[2] 山本博史, 匂坂芳典: “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 年度秋期研究発表会講演論文集, I,2-1-19,pp.75-76 Sep.1998

[3] 山本博史: "多重クラスN-gramによる効率的言語モデル表現" 日本音響学会平成10年度秋期研究発表会講演論文集, I,2-1-20,pp.77-78 Sep.1998

[4] Christer Samuelsson , Wolfgang Reichl "A Class-based Language Model for Large-vocabulary Speech Recognition extracted from Part-of-speech Statistics" IEEE Proc. ICASSP'99 ,vol.1 , pp537-540 , 1999

[5] 政瀧浩和, 松永昭一, 匂坂芳典 "連続音声認識のための品詞・単語可変長 N-gram" 音講論, 1-P-17, pp.195-196, Mar.1996.

[6] F.Jelinek, R.L.Mercer. "Interpolated Estimation of Markov Source Parameters from Sparse Data". In Pattern Recognition in Practice, pp.381-397.North Holand,1980

[7] T.R.Niesler,P.C.Woodland. "Combination Of Word-Based And Category-Based Language Models". In Procs.ICSLP'96,pp.1779-1782,1997

[8] A.Farhat,J.Isabelle,D.O'Shaughnessy "Clustering Words for Statistical Language Models Based on Contextual Word Similarity" IEEE Proc. ICASSP'96, vol.1 , pp.180-183 , 1996

[9] T.R.Niesler,P.C.Woodland "A Variable-Length Category-Based N-gram Language Model" IEEE Proc. ICASSP'96, vol.1 , pp.164-167 , 1996 W.Ward,S.Issar: "A Class Based Language Model for Speech Recognition" IEEE Proc. ICASSP'96 , vol.1 , pp416-418 , 1996

[10] Michèle Jardino , Gilles Adda "Automatic Word Classification Using Simulated Annealing" IEEE Proc.ICASSP'93,Vol.1, pp.41-44 ,1993

[11] Michèle Jardino "Multilingual Stochastic N-gram Class Language Models" IEEE Proc.ICASSP'96,Vol.1,pp.161-163 ,1996

[12] "毎日新聞 CD-ROM 1994年版" 毎日新聞社 ,1995

[13] "新聞記事読み上げ音声コーパス" JNAS Vols.1-16 , 日本音響学会 ,1997