音声言語情報処理 34－39
(2000. 12. 22)

# 連語を組み込んだ統計言語モデル

岩瀬 修　　森元 逞　　首藤 公昭
福岡大学大学院 工学研究科 電子工学専攻
〒814-0180　福岡市城南区七隈8丁目19番1号
電話番号:(092)871-6631(内線:6572)
**E-mail:** iwase@mdmail.tl.fukuoka-u.ac.jp, {morimoto, shudo}@tl.fukuoka-u.ac.jp

あらまし

日本語には慣用表現などの比較的固定的な表現 (以下、「連語」と呼ぶ) が数多くある。我々は、この連語を従来の統計言語モデルに組み込む方法を提案し、またその効果について検証する。テストセットパープレキシティを求めた結果、従来の言語モデル (バイグラム) と比較して、連語が出現する評価データでは約8％の性能が改善できた。しかし、連語が出現しない評価データでは、若干の性能悪化が見られた。そこで、この性能悪化の問題の改善方法についても考察する。

キーワード　連語、統計言語モデル、テストセットパープレキシティ

# Incorporating Collocation Date into a Statistical Language Model

Osamu Iwase　　Tsuyoshi Morimoto　　Kosho Shudo
Graduate School of Engineering, Fukuoka University
8-19-1 Nanakuma, Jounan-ku, Fukuoka, 814-0810 Japan
**TEL:** (092)871-6631(ext. 6572)
**E-mail:** iwase@mdmail.tl.fukuoka-u.ac.jp, {morimoto, shudo}@tl.fukuoka-u.ac.jp

Abstract

In Japanese, there are many collocational expressions such as idiomatic wording. We already collected thousands of Japanese collocation date. In this paper, a method incorporating the collocation date into a statistical language model is proposed, and the evaluation results are reported. Test-set-perplexity is improved about 8% when applied to sentences which include collocation expressions. However, when applied to sentences which do not include such expressions, the perplexity become slightly worse. Some improving methods for coping with the problem are also mentioned.

key words　collocation date, statistical language model, test-set-perplexity

## 1 はじめに

近年、大語彙(数千～数万語)を対象とした音声認識システムが開発されている。音声認識の精度は、音素モデルの良否だけでなく、言語モデルの良否にも大きく依存する。

言語モデルの中で、最も一般的に用いられるものは、新聞記事などの多量の学習データから統計的な処理によって抽出されたNグラムである。このNグラムは、Nの値が大きい方が望ましいが、実際は学習データの量が限られているため、N＝2(バイグラム)、N＝3(トライグラム)が用いられることが多い。

一方、日本語には、かなり長い単語列からなる慣用的・半固定的な表現が多く存在する。我々は、このような表現(以下、「連語」と呼ぶ)を新聞、雑誌、小・中学校の教科書、各種辞書などから、人手で採集した[1][2]。この連語の例を表1に示す。

表1 連語データの例(-:形態素の区切りの印)

| 形態素数 | 個数 | 連語の例 |
|---|---|---|
| 2 | 11,188 | 明らか-に、一緒-に、消極-的 |
| 3 | 21,532 | 涙-が-出る、舞台-に-上がる |
| 4 | 6,332 | 首-を-長く-する |
| 5 | 2,860 | 目-から-火-が-出る |
| 6 | 940 | 猫-の-手-も-借り-たい |
| 7 | 537 | 他人-の-出る-幕-で-は-ない |
| 8 | 219 | 二-度-ある-こと-は-三-度-ある |
| 9 | 119 | 縦-の-もの-を-横-に-も-し-ない |
| 10以上 | 124 | 右-を-見-て-も-左-を-見-て-も |

計：43,851個

このような連語を言語モデルに組み込むことができれば、言語モデルの性能を大幅に改善することができると思われる。しかし、連語の収集に当たって網羅性を第一にしたため、残念ながら数量に関する情報(統計値など)は得られていない。

本研究では、ある統計言語モデルが与えられた時(以下、この言語モデルを「ベースの言語モデル」と呼ぶ)、その言語モデルに上記の連語データを組み込む方法について述べる。また、その効果について述べる。

なお、具体的なベースの言語モデルとしては、情報処理振興事業協会(IPA)から研究用に提供されている音声認識システムJulius[3]の言語モデルを用いた。このJuliusの言語モデルには、「バイグラム」と「逆向きトライグラム」があるが、今回の研究では、まず第一段階としてバイグラムを用い、これに3単語以上の連語を組み込むことにした。

## 2 連語の検証

連語は慣用的・半固定的な表現であるので、それらを構成する単語相互の連接確率は大きいと予想される(逆に言えば、このような連接確率が大きいものが連語として採用されるべきである)。そこで、我々が採取した連語データがこのように連接確率の観点から妥当なものであるかどうかを、ベースの言語モデルに定義されている確率値を用いて検証することにした。ただし、ベースの言語モデル内に定義されているのはバイグラム確率であるから、検証の対象としたのは、2単語連語のみである。

具体的には、2単語連語に対応するバイグラム確率の対数を求め、その分布を求めた。また、2単語連語以外の2単語列についても同様な分布を求め、両者の分布を比較した。語彙数5,000語、カットオフが0の場合の結果を図1に示す。

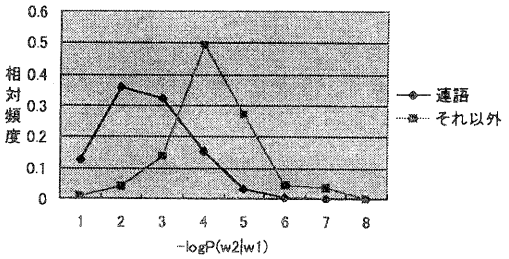


図1 2単語連語と連語以外のバイグラムの比較
(語彙数5,000語、カットオフ0)

横軸は対数尤度にマイナスを付けたものであるから、尤度値が小さい方が確率が高いことになる。縦軸は相対頻度である。なお、語彙数やカットオフ条件が異なっても傾向はほぼ同じである。

この図から、2単語連語を構成する単語相互の連接確率は、それ以外のものよりかなり大きい値を持つものが多いことがわかる。ただし、2単語連語は全部で約11,000個あるが、上記で対象とした連語はJulius内のバイグラムにエントリが存在した約1,800個のみである。このようにエントリに存在しなかった連語が多い理由は、恐らく連語データが極めて広範な分野から抽出されたためであると考えられる。しかし、少なくともJuliusが対象としている言語モデルの分野に関しては、連語データは十分有効であると考えられる。表2に、連接確率値が大きなもの上位10個を示している。

表2 2単語連語で確率(対数尤度)の大きなもの
(上位10個)

| 順位 | 確率 | 2単語連語 |
|---|---|---|
| 1 | -0.0111 | いけ-ない |
| 2 | -0.0191 | 従業-員 |
| 3 | -0.0476 | 明らか-に |
| 4 | -0.0549 | 一緒-に |
| 5 | -0.0572 | 多角-的 |
| 6 | -0.0588 | 特捜-部 |
| 7 | -0.0654 | 評-議会 |
| 8 | -0.0700 | 消極-的 |
| 9 | -0.0957 | 対照-的 |
| 10 | -0.1070 | 同時-に |

# 3 連語N-gram

## 3.1 連語N-gramの推定方法

連語と統計言語モデルを組み合わせるには様々な方法が考えられる。例えば、ある単語列Wの尤度$S_t$を統計言語モデルから得られる尤度$S_s$と、連語としての尤度$S_c$の線形結合で求める方法などがある。しかしこの方法では$S_s$と$S_c$の重みをどう設定するかなどの問題がある。

そこで、我々は、連語をNグラムの確率モデルとしてモデル化することを考える。この方法は、全体を統計言語モデルの枠組みで統一的に取り扱うことができるため、見通しが良い。以下で、その連語N-gramの求め方を述べる。

一般に、ある単語列W=$w_1w_2...w_N$に対する生起確率は、以下のように求めることができる。

$$P(W) = P(w_1) \times P(w_2|\, w_1) \times P(w_3|\, w_1w_2) \times ... \times P(w_N|\, w_1w_2...w_{N-1}) \quad \cdots (1)$$

この式の第3項以降は、いわゆるNグラム(N $\geq$ 3)であるが、連語に関しては、

$$P(w_k|\, w_1...w_{k-1}) \gg P(w_k|\, w_{k-1}), \quad 3 < k < N$$

になると考えられるから(図2)、バイグラムの近似では不十分であろう。以下では、このNグラムの求め方について述べる。

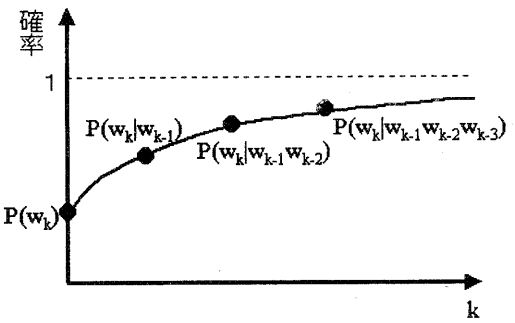


図2 連語N-gram

加藤ら[4]は放送ニュース文を対象としたNグラム言語モデルを提案している。そこでは、Nグラムに基づく単語の生起確率$P$は$P = P_0 \times (1 - e^{-\lambda N})$のようにNが大きくなるにつれて、漸近的に一定値$P_0$に近づくとしている。我々も基本的にはこの考えを採用している。しかし、加藤らは、$P_0$や$\lambda$は単語によらず一定値としているが、我々はこれらの値をベースの言語モデルから求めることにした。すなわち、以下の式が成り立つと仮定する。

$$P(w_k|\, w_1w_2...w_{k-1}) = P(w_k) + (1 - P(w_k)) \times (1 - e^{-\lambda_k(k-1)}) \quad 2 \leq k \leq N \quad \cdots (2)$$

ここで、$\lambda_k$はベースの言語モデルのユニグラム、バイグラムから求める。すなわち、(2)式をバイグラムに適用すると、

$$P(w_k|\, w_{k-1}) = P(w_k) + (1 - P(w_k)) \times (1 - e^{-\lambda_k}) \quad \cdots (3)$$

が成り立つから、これより$\lambda_k$は、

$$\lambda_k = -\log \frac{1 - P(w_k|\, w_{k-1})}{1 - P(w_k)} \quad \cdots (4)$$

として、求めることができる。このようにして、連語W=$w_1w_2...w_N$の各単語に対する$\lambda_k$を求めた後、連語

W 全体に対する λ を各 $\lambda_k$(k=2,3,...,N) の平均値として求める。

$$\lambda = (\lambda_2 + \cdots + \lambda_N)/(N-1) = -\frac{1}{N-1}\sum_{k=2}^{N}\log\frac{1-P(w_k|\,w_{k-1})}{1-P(w_k)} \quad \cdots (5)$$

### 3.2 組み込む連語の基準

上の方法により、各連語データに対する λ を求めた。3 単語以上の連語 32,663 個のうち、バイグラムデータが存在したものが 3,906 個であった (存在したデータが少ない理由は 2 章で述べた)。求めた λ のうち、プラスになったものが 3,375 個、マイナスになったものが 531 個という結果になった (表 3 )。

表 3 λ の内訳

| λ の値の正負 | + | − |
|---|---|---|
| 連語の個数 | 3,375 | 531 |

計 3,906 個

まず、マイナスのものは図 2 のようにならない。従って、これらの連語は言語モデルに組み込むと性能を下げてしまうおそれがあるので、言語モデルに組み込まない。次に、λ がプラスの場合であるが、上述したように λ は $\lambda_k$(k=2,3,...,N) の平均値としている。このため、連語の中には推定した生起確率の方がバイグラムの積で求めた生起確率よりも小さくなることも考えられる。そこで、両者の比較を行ない、推定した確率がバイグラムによる確率に比べて大きくなった連語 (表 6 ) のみを言語モデルに組み込む (表 4 )。

表 4 組み込む連語の内訳

| [a] 推定生起確率と [b] バイグラムによる生起確率の差 | [a]-[b] > 0 | [a]-[b] < 0 |
|---|---|---|
| 連語の個数 | 3,054 | 321 |

表 5 生起確率が大きな連語 (上位 1 0 個)
(確率の欄は確率値の対数尤度)

| 順位 | 生起確率 | 連語 |
|---|---|---|
| 1 | -3.4988 | それ-に-も-かかわら-ず |
| 2 | -3.5264 | これ-を-もっ-て |
| 3 | -3.5993 | か-と-いっ-て |
| 4 | -3.6036 | 一-呼吸-置い-て |
| 5 | -3.6194 | 何-と-なれ-ば |
| 6 | -3.6447 | 日-を-追っ-て |
| 7 | -3.6673 | 一-人-の |
| 8 | -3.6920 | 明らか-に-なる |
| 9 | -3.6952 | 明らか-に-する |
| 10 | -3.7055 | ある-種-の |

表 6 バイグラムによる生起確率と推定生起確率の差
(上位 1 0 個)
(確率の欄は確率値の対数尤度)

| 順位 | 推定した生起確率 | バイグラムによる生起確率 | 連語 |
|---|---|---|---|
| 1 | -10.5326 | -20.3417 | 仏-の-光-より-金-の-光 |
| 2 | -9.5874 | -19.3516 | 苦労-を-苦労-とも-思わ-ない |
| 3 | -12.0355 | -21.6704 | 目-で-見-て-口-で-言え |
| 4 | -10.2019 | -19.7705 | 海-とも-山-とも-つか-ず |
| 5 | -5.6375 | -15.1977 | 面-と-向かっ-て-の |
| 6 | -7.6028 | -17.1392 | 後-の-世-まで-伝える |
| 7 | -5.7031 | -15.0940 | 犠牲-者-が-出る |
| 8 | -7.9759 | -17.2969 | 一-人-相撲-を-取る |
| 9 | -4.6985 | -13.8546 | 万-に-一つ-の-可能-性 |
| 10 | -6.2041 | -15.3498 | 首-を-長く-し-て-待つ |

## 4 バイグラム言語モデルへの連語の組み込み

前章で、連語 N グラムの生起確率の計算方法について述べた。しかし、このようにして求めた N グラムをそのままの N グラムとして音声認識システム (Julius) に組み込むには、システムの探索アルゴリズム[1] を変える必要がある。

そこで、N 単語からなる連語 $W=w_1w_2...w_N$ において $W$ を 1 語と考え、そのバイグラム、すなわち、P(W | a) および、P(b | W) を求め (a,b はそれぞれ連語 $W$ に前接

[1] Julius は 2 パス探索を採用しており、第一パスでバイグラム、第二パスで逆向きトライグラムを用いている。

および後接する単語)、これを従来のバイグラム言語モデルに新たに組み込むことにする。

**(i) P(W | a) について**

まず前接のバイグラム確率 P(W | a) について述べる。W を $w_1 w_2 ... w_N$ に展開し、それを変形すると、

$$\begin{aligned} P(W \mid a) &= P(w_1 w_2 ... w_N \mid a) \\ &= P(w_1 \mid a) P(w_2 ... w_N \mid w_1) \\ &= \{P(w_1 \mid a)/P(w_1)\} P(w_1 w_2 ... w_N) \end{aligned} \quad \cdots (6)$$

となる。この式において、$P(w_1 w_2 ... w_N)$ は前述の方法により求める。また、P($w_1$ | a) および P($w_1$) は、従来の言語モデルで与えられているものを用いる。

一方、これを既存のバイグラム言語モデルに組み込む際には、この P(W | a) を他のバイグラムの確率値から差し引く必要がある。そこで、連語の先頭単語を $w_1$ とすると、ある単語 (ここでは a) の後に $w_1$ がくるバイグラム確率値 P($w_1$ | a) から以下の式 (7) のように組み込もうとする連語の確率値分を差し引く。

$$\hat{P}(w_1 \mid a) = P(w_1 \mid a) - \sum P(W \mid a) \quad \cdots (7)$$

**(ii) P(b | W) について**

次に、連語 W を 1 語とみなした場合のある単語が後接する確率値 P(b | W) であるが、これは単純に連語 W の最後の単語 $w_N$ のバイグラムを用いて、以下のように求めることにする。

$$P(b \mid W) = P(b \mid w_N) \quad \cdots (8)$$

## 5 実験と評価

前章で、述べた方法により、連語を組み込んだバイグラム言語モデルを構築し、この言語モデルの性能をテストセットパープレキシティにより評価した。この計算には CMU-Cambridge Toolkit[5] を利用した。

評価用のテキストには毎日新聞を用いて、**(a) 連語が出現する文 1 万文、(b) 連語の出現が見られない文 1 万文**、でそれぞれ従来のバイグラム言語モデルと比較を行なった (表 7 )。

表 7 従来の言語モデルと連語を組み込んだ言語モデルの性能比較 (バイグラム)

| | (a) 連語の出現あり | (b) 連語の出現なし |
|---|---|---|
| [1] 従来のバイグラム言語モデル | 62.99 | 64.22 |
| [2] 連語を組み込んだ言語モデル | 58.48 | 65.09 |
| パープレキシティの差 ([1]-[2]) および比率 (↑:改善↓:悪化) | 4.51 (8 % ↑) | -0.87 (1 % ↓) |

※一文当たり：平均 35 単語

連語を一単語として従来のバイグラム言語モデルに組み込むことにより、**(a) 連語が出現する文**では約 8 %のパープレキシティの減少が確認された。しかし、**(b) 連語の出現が見られない文**では 1 %弱のパープレキシティの増加が見られた。この原因として、以下の事が考えられる。

1) 連語を組み込むことによるエントロピーの増加
2) 評価データ中においての連語の取りこぼし

1 ) については, 組み込む連語の確率値を従来のバイグラムから差し引いたことが、悪影響を及ぼしていると考えられる。この問題を解決する方法として、連語のうち、生起確率の高いものだけを組み込むことが考えられる。たとえば、$\lambda$ にある閾値 $\Lambda$ を設けて、$\lambda \geq \Lambda$ となる連語のみを組み込む対象にする。

2 ) については, 連語を一単語として言語モデルに組み込んでいるため、評価データ中の連語の表記・表現の揺れに対処できないためと考えられる (表 8 )。

表 8 評価データ中の連語の表記・表現の揺れ

| 連語 | 表記・表現の揺れ |
|---|---|
| 影-も-形-も-見え-ない | 影-も-形-も-見当たら-ない |
| 額-に-汗-し-て | 額-に-汗-を-にじま-せ-ながら |
| 身-に-覚え-が-ある | 身-に-覚え-の-ある |
| ここ-ぞ-と-いう-時 | ここ-ぞ-と-いう-場面 |
| 今-に-して-思う-と | 今-に-して-思え-ば |
| 面-と-向かっ-て-の | 面-と-向かう、面-と-向かっ-て |
| 犠牲-者-が-出る | 犠牲-者-が-でる |

この問題を解決する方法として、連語 $W = w_1 w_2 ... w_N$ のプレフィックス $W_1^i = w_1 w_2 ... w_i (3 \leq i \leq N-1)$ すべ

てを取り出し、これを、バイグラム言語モデルに組み込むことが考えられる。

## 6 おわりに

日本語における慣用表現などの比較的固定的な表現(連語)を従来の統計言語モデルに組み込む方法を提案し、その効果を検討した。連語を従来のバイグラム言語モデルに組み込むことにより従来の言語モデルと比較して、連語が出現する文ではパープレキシティの減少が確認された。

しかし、連語の出現しない文に対しては従来の言語モデルに比べてパープレキシティが悪くなる結果になった。これを改善する方法として、(1)**生起確率の高い連語のみを組み込む**、(2)**連語のプレフィックスについてもそれらを連語として組み込む**、ことが考えられる。今後は、(1)と(2)の種々の組合せについて実験を進め、その最適な組合せを明らかにしたい。また、連語のプレフィックスをすべて連語とする単純な方法だけでなく、表現・表記間の距離尺度を用いることにより揺れを吸収する方法についても検討を進めていきたい。

**謝辞**

本研究を進めるにあたり、財団法人サウンド技術振興財団の研究助成金を頂いた。また、情報処理振興事業協会の Julius、CD-ROM 版毎日新聞記事データを使用させて頂いた。ここに謝意を表したい。

## 参考文献

[1] 安武, 小山, 吉村, 首藤 :「固定的共起表現とその変化形」, 言語処理学会第3回年次大会発表論文集, pp.449-452, 1997.3

[2] 小山, 安武, 吉村, 首藤 :「連語データを利用した仮名漢字変換」, 情報処理学会論文誌, Vol.39, No.11, pp.2978-2987, 1998.11

[3] 河原, 李, 小林, 武田, 峯松, 伊藤, 山本, 山田, 宇津呂, 鹿野 :「日本語ディクテーション基本ソフトウェア(97年度版)の性能評価」, 情報処理学会研究報告, 98-SLP-21-10, 1998

[4] 加藤, 浦谷, 江原, 安藤 :「ニュース音声認識のための(n ≥ 4)-gramを併用する言語モデル」, 情報処理学会研究報告, 99-SLP-29, pp.187-192, 1999

[5] Clarkson, P. and Rosenfeld, R. : Statistical Language Modeling Using the CMU-Cambridge Toolkit, Proc. Eurospeech'97, 1997