音声言語情報処理 34－38
(2000. 12. 22)

多重クラス Trigram 構築のための効率的な自動クラスタリング手法

磯貝俊太郎† 白井 克彦† 山本 博史†† 匂坂 芳典††

† 早稲田大学理工学部
〒 169–8555 東京都新宿区大久保 3–4–1
†† ATR 音声言語通信研究所
〒 619–0288 京都府相楽郡精華町光台 2–2

E-mail: †{isogai,shirai}@shirai.info.waseda.ac.jp, ††{hirofumi.yamamoto,yoshinori.sagisaka}@slt.atr.co.jp

あらまし 本稿は，多重クラス trigram のための効率的な自動クラスタリング手法を提案する．従来手法のように単語履歴をクラスタリングするのではなく，'DUAME Language Modeling を用いた単語 trigram の近似手法' を自動クラスタリング手法へ適応することにより，効率的に自動クラスタリングを行うことが出来る．本手法で分類したクラスを基に多重クラス trigram を構築した結果，単語 trigram の 100 分の 1 以下のパラメータサイズであるにもかかわらず，パープレキシティ・単語認識率による評価で共に単語 trigram を上回る性能を示した．

キーワード クラス N-gram, 多重クラス N-gram, 自動クラス分類, DUAME Language Modeling

The efficient method of automatic clustering for Multi-Class Trigrams

Shuntaro ISOGAI†, Katsuhiko SHIRAI†, Hirofumi YAMAMOTO††, and Yoshinori SAGISAKA††

† School of Science and Engineering, Waseda University
Okubo 3–4–1, Shinjuku-ku, Tokyo, 169–8555 Japan
†† ATR Spoken Language Translation Research Laboratories
Hikaridai 2–2–2, Seika-cho, Soraku-gun, Kyoto 619–0288 Japan

E-mail: †{isogai,shirai}@shirai.info.waseda.ac.jp, ††{hirofumi.yamamoto,yoshinori.sagisaka}@slt.atr.co.jp

Abstract In this paper , an efficient automatic word clustering method is proposed for Multi-Class Trigrams . The third position words in the trigrams are directly clustered using 'word trigram approximation by DUAME Language Modeling '. Therefore, conventional word-history clustering is not required. The Multi-Class Trigrams based on these classes showed better performance both in perplexity and recognition rates compared to conventional word trigrams. Additionally the parameter size can be reduced down to one percent.

Key words Class N-gram, Autimatic Clustering, Multi-Class N-gram, DUAME Language Modeling

1. はじめに

大語彙連続音声認識の言語モデルとしては，単語 N-gram が広く用いられている．単語 N-gram は N を大きくすれば次単語への予測精度の向上が期待できるが，N の増加に伴って単語遷移の組合せの数も急激に増加し，モデルサイズが増大する．また，大量の学習データを必要とし，観測データの少ない単語ペアに対しては統計的に信頼できる確率値が得られないという問題がある．

これらの問題の解決手段として，クラス N-gram が提案されている．クラス N-gram は，複数の単語をまとめてクラスとし，単語間の遷移をクラス間の遷移で近似するモデルである．さらに，クラス N-gram を改良したものとして，「接続の方向性を考慮した多重クラス N-gram モデル」が山本らによって提案されている [1]．多重クラス N-gram モデルは，クラス N-gram において単語間の接続性を表わすクラスを，先行・後続の二つの方向に分けて考えるモデルである．文献 [1] によると，多重クラス 2-gram はパープレキシティ・連続単語認識による評価でクラス 2-gram・単語 2-gram を上回る性能を示している．単語 N-gram と同様に，多重クラス N-gram においても，N を大きくすれば性能の向上が期待できるため，多重クラス 3-gram を構築すれば，より良い性能を得られると考えられる．

その際に問題となるのが，クラスタリング手法である．従来の様に一つ前・二つ前の単語のような単語の履歴をクラスタリングすることを考えると，N が増えるに従い単語履歴数が大幅に増加するので，計算量も大幅に増えることになる．そのため，クラスタリングを行うのに非常に時間がかかる．

そこで本稿では，多重クラス 3-gram のためのクラスタリングを効率的に行う手法を提案する．また，多重クラス 3-gram を構築し，その性能評価を行う．

2. 多重クラス N-gram

2.1 クラス N-gram

単語 N-gram は直前の $N-1$ 個の単語列から次単語の出現確率を統計的に予測するモデルであり，出現確率は (1) 式の様に与えられる．

$$P(w_i|w_{i-N+1},\cdots,w_{i-2},w_{i-1}) \tag{1}$$

そして，単語 N-gram における問題を解決するための方法としてクラス N-gram が提案されている．クラス N-gram においては，直前の $N-1$ 個の単語列から次単語の出現確率は (2) 式の様に与えられる．

$$P(w_i|c_i)P(c_i|c_{i-N+1},\cdots,c_{i-2},c_{i-1}) \tag{2}$$

ここで，c_n は単語 w_n が属するクラスを表す．

2.2 多重クラス N-gram

クラス N-gram において，単語の前方向・後方向の接続性を別の属性としてみなし，各単語にその属性ごとに複数のクラスを割り当てる多重クラス N-gram が山本らによって提案されている [1]．先の (2) 式を多重クラス N-gram にあてはめると，(3) 式のように表される．

$$P(w_i|c_i^t)P(c_i^t|c_{i-N+1}^{fN-1},\cdots,c_{i-2}^{f2},c_{i-1}^{f1}) \tag{3}$$

ここで c_n^{fm} は条件部分の m 番目に現れた場合に単語 w_{n+1} に割り当てられるクラスを表す．また，c^f は先行する場合のクラス (以下，from クラスと呼ぶ)，c^t は後続する場合のクラス (以下 to クラスと呼ぶ) を表す．

$N=2$ とした多重クラス 2-gram については，パープレキシティ・連続単語認識による評価で従来のクラス 2-gram・単語 2-gram を上回る性能を示したことが報告されている [1]．

本稿で扱う多重クラス 3-gram は，(3) 式において $N=3$ の場合，つまり

$$P(w_3|c_3^t)P(c_3^t|c_1^{f2},c_2^{f1}) \tag{4}$$

と表される．(4) 式に出現するクラスを効率的にクラスタリングする手法を提案することが，本稿の目的である．

3. 自動クラスタリング

3.1 自動クラスタリングの目的

クラスタリングの基準としては様々なものが考えられる．その中でも，品詞に基づいたクラスタリングがよく用いられる．品詞情報に基づく方法は，学習データに出現しない単語に対してもクラスを割り当てられるという利点がある．しかし，品詞情報にのみ基づいたクラスタリングは，N-gram モデルにとって最も重要である単語間の統計的な接続特性を詳細に表わしているとは言い難い．そこで，本稿では品詞情報に加え，単語の接続性に注目した自動クラスタリングを行った．

3.2 単語履歴のクラスタリング

従来のように単語履歴のクラスタリングを行うのであれば [6]，多重クラス 3-gram のためのクラスタリングには，図 1 のように，単語 W_{n-2} と W_{n-1} の単語ペア W' と，単語 W_n との間でクラスタリングを行うことになる．そうなると、単語ペア W' の数

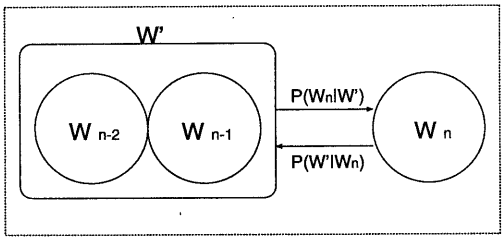

図1 単語履歴のクラスタリング
Fig. 1 word-history clustering

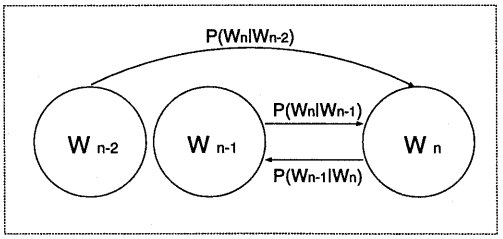

図3 'Distance-2bigram を用いた単語 trigram の近似手法' のクラスタリングへの適応
Fig. 3 Clustering with 'word 3-gram approximation by Distance2-bigram'

が認識対象単語数の自乗となるため，パラメータ数が膨大な量になり、計算量が非常に多くなる．そこで，パラメータ数削減のために，3.3のような近似手法を用いる．

3.3 DUAME Language Modeling [2] を用いた単語 3-gram の近似手法

従来の単語N-gramを低次の単語N-gramで近似するための手法として、'Distance -related Unit Association Maximum Entropy(DUAME) Language Modeling' が提案されている [2]．文献 [2] より，図2のように単語 2-gram($P(w_n|w_{n-1})$) と，Distance2-bigram($P(w_n|w_{n-2})$) を最大エントロピー法で統合することにより，単語 3-gram($P(w_n|w_{n-2}^{n-1})$) の良い近似が得られることが分かっている．これを，クラスタリングの際の良い指標として用いることが出来ると考え，3.4 の様にクラスタリングを行う．

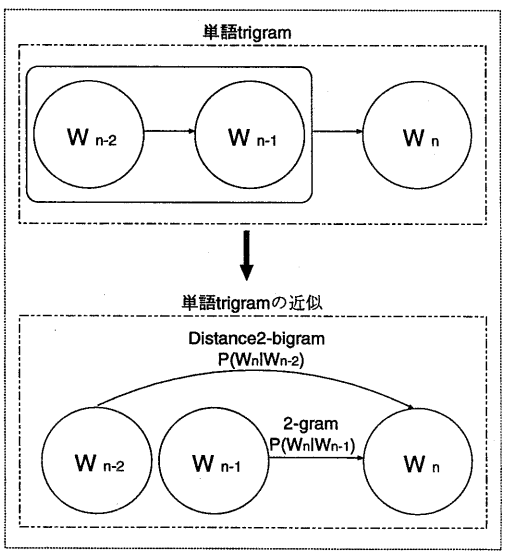

図2 Distance2-bigram を用いた単語 3-gram の近似
Fig. 2 Approximation of word 3-gram by Distance2-bigram

3.4 'Distance2-bigram を用いた単語 3-gram の近似手法' のクラスタリングへの適応

3.3 から，単語 W_{n-2} のクラスタリングの際に，単語 W_n との関係だけを考慮するだけで，クラス C_{n-2} を良い精度でクラスタリング出来ると考えられる．そこで，この Distance2-bigram を用いた単語 3-gram の近似手法をクラスタリングに適応し，図3のように3つの 2-gram(前向き 2-gram・後向き 2-gram・Distance2-bigram) を使ってクラスタリングを行う．図3のように Distance2-bigram を使って単語 W_{n-2} をクラスタリングすることによりクラス C_{n-2} を得る．

3.5 自動クラスタリング手法

自動クラスタリング手法としては，[3] [4] をはじめ，いくつかの提案がされているが，今回は，山本 [1] による手法を用いた．

クラスタリングの手順は以下の通りで，分類の対象としたのは，to クラス・from クラス・Distance2-from クラスである．

（1） 初期クラスとして，1単語1クラスとする．

（2） 個々の単語またはクラス X に対して，次のようなベクトルを与える．

$$V_t(X) = \{P_t(W_1|X), P_t(W_2|X), \cdots, P_t(W_n|X)\}$$
$$V_f(X) = \{P_f(W_1|X), P_f(W_2|X), \cdots, P_f(W_n|X)\}$$
$$V_{fd2}(X) = \{P_{fd2}(W_1|X), P_{fd2}(W_2|X), \cdots, P_{fd2}(W_n|X)\}$$

ここで，$P_t(W_i|X)$, $P_f(W_i|X)$ は，単語またはクラス X から単語 W への後向き及び前向きの bigram の確率値を表す．$P_{fd2}(W_i|X)$ は，単語またはクラス X から単語 W への前向きの Distance2-bigram 確率値を表す．

（3） U_{new}, U_{old} を次のように定義し，マージコスト $U_{new} - U_{old}$ が最小となるようなクラスのペアを選び，統合して一つのクラスとする．

$$U_{new} = \sum_{w} P(W) \times D(V(C_{new}(W)), V(W))$$

$$U_{old} = \sum_{w} P(W) \times D(V(C_{old}(W)), V(W))$$

ここで，C_{new} は統合後のクラス，C_{old} は統合前のクラスを表す．$D(V_C, V_W)$ はベクトル V_C と V_W のユークリッド距離の自乗を表す．

（4） (2),(3) の手順をあらかじめ定められたクラス数になるまで繰り返す．

4. 多重クラス trigram の性能評価

4.1 パープレキシティによる性能評価

前節でクラスタリングした3種類のクラスを用いて多重クラス 3-gram を構築し，パープレキシティによる評価を行った．

訓練セットは総単語数 1606951・異り単語数 16355，評価は訓練セットに含まれない対話データ 42 片対話 6326 語の評価セットを使用した．

まず，予備実験として，多重クラス 2-gram においてパープレキシティ最小となるクラス数を調べるための実験を行った．図 4は，単語 2-gram と多重クラス 2-gram について，多重クラス 2-gram のクラス数を横軸，パープレキシティを縦軸にとったものである．多重クラス 2-gram の to クラス数 (C_{n-1})・from クラス数 (C_n) を最適化することは本実験の目的ではないので，to クラス数と from クラス数は同数とした．図 4において，クラス数 1000 で多重クラス 2-gram(図 4中 Multi-Class-2gram) のパープレキシティが最小となっている．このクラス数を，次の実験でのクラス C_n,C_{n-1} の数として固定して使用している．

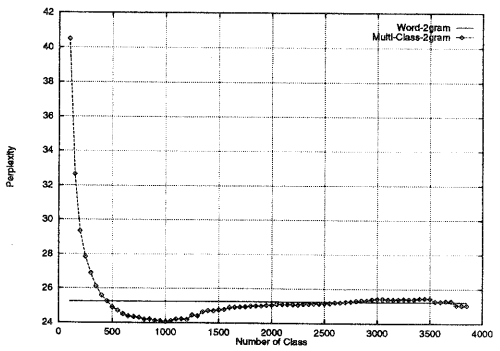

図 4 多重クラス 2-gram と単語 2-gram のパープレキシティによる比較

Fig. 4 Conparison between Multi-Class 2-gram and word 2-gram with perplexity

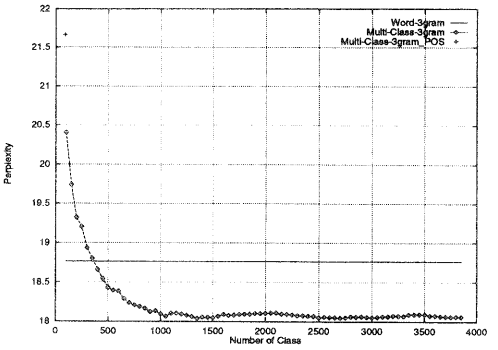

図 5 クラス C_{n-2} の違いによるパープレキシティの比較

Fig. 5 Conparison between models' perplexity with Class-C_{n-2} difference

図 5はそれぞれ，

- 単語 3-gram(図 5中 Word-3gram)
- Distance2-bigram を用いてクラス C_{n-2} をクラスタリングした「多重クラス 3-gram(提案手法)」(図 5中 Multi-Class-3gram)
- クラス C_{n-2} に品詞クラス (89 クラス) をとった「品詞-多重クラス 3-gram」(図 5中 Multi-Class-3gram_POS)

についてそれぞれパープレキシティ(縦軸) を示したものである．横軸は多重クラス 3-gram のクラス C_{n-2} の数をとっている．

図 5より，今回提案した多重クラス 3-gram が，単語 3-gram・品詞-多重クラス 3-gram よりも低いパープレキシティを示している．

この結果より，単語 3-gram における学習データの不足が原因の統計的信頼性の低さを，多重クラス 3-gram ではうまく解消出来たと考えられる．また，品詞情報はコーパスに現れない単語に対してもクラスを割り当てることができるが，品詞のみに基づくクラスよりも，実際のコーパスから単語の接続性を統計的に反映させたほうが，より良いクラスを得ることができたと言える．

次に，表 1は各モデルの論理パラメータ数と先のパープレキシティを示したものである．表 1より，多重クラス 3-gram(表 1 中 MC 3-gram) のクラス数 500 と単語 3-gram を比較すると，論理パラメータ数はの 100 分の 1 であるのにもかかわらず，パープレキ

表1 パープレキシティによる性能評価
Table 1 Performance evaluation with perplexity

モデル (クラス数)	論理パラメータ数	Perplexity
MC 3-gram (500)	5.0×10^8	18.60
MC 3-gram (700)	7.0×10^8	18.39
MC 3-gram (1000)	1.0×10^9	18.25
MC 2-gram (500)	2.5×10^5	24.89
MC 2-gram (700)	4.9×10^5	24.34
MC 2-gram (1000)	1.0×10^6	24.06
単語 2-gram	1.5×10^7	25.23
単語 3-gram	5.6×10^{10}	18.76

シティは多重クラス 3-gram が上回っていることが分かる．つまり，多重クラス 3-gram が小さなサイズで高い性能を示していることになる．モデルサイズが小さいということは，大語彙連続音声認識における統計的言語モデルにおいて，非常に重要である．

4.2 連続単語認識による性能評価

表2は，パープレキシティによる評価と同様に、訓練セットに含まれない対話データ 42 片対話 6326 語の評価セットを使用した連続単語認識実験の単語正解精度である．正解精度の算出には (5) 式を用いた．

$$Accuracy = \frac{W - D - I - S}{W} \tag{5}$$

ここで，

- W : 正解単語総数
- D : 脱落誤り数
- I : 挿入誤り数
- S : 置換誤り数

である．

単語認識精度においても，単語 3-gram の 100 分の 1 程度の論理パラメータ数である多重クラス 3-gram が，単語 3-gram を上回る性能を示した．

5. ま と め

本稿では，Distance2-bigram を用いてクラス C_{n-2} のクラスタリングを行い，さらにそのクラスを用いて多重クラス 3-gram を構築し，その評価を行った．

前節での実験結果の通り，多重クラス 3-gram はパープレキシティ・単語正解精度による評価で，少ない論理パラメータ数であるにもかかわらず単語 3-gram を上回る性能を示し，本手法の有効性を確かめることが出来た．

また，本稿では，クラス C_n, C_{n-1} の数を最適化していないため，論理パラメータ数をさらに減少させることが可能である．

表2 連続単語認識による性能評価
Table 2 Performance evaluation with recognition

モデル (クラス数)	論理パラメータ数	Accuracy(%)
MC 3-gram (500)	5.0×10^8	86.81
MC 3-gram (700)	7.0×10^8	87.39
MC 3-gram (1000)	1.0×10^9	87.56
MC 2-gram (500)	2.5×10^5	85.75
MC 2-gram (700)	4.9×10^5	86.17
MC 2-gram (1000)	1.0×10^6	86.11
単語 2-gram	1.5×10^7	84.45
単語 3-gram	5.6×10^{10}	86.05

文 献

[1] 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 年秋期研究発表会論文集.

[2] Shuwu. Zhang, Harald. Singer, Dekai. Wu, Yoshinori. Sagisaka,"Improving N-gram Modeling using Distance-Related Unit Association Maximum Entropy Language Modeling", Proceedings of Eurospeech'99, Vol. 4, pp. 1611-1614, 1999

[3] Shuanghu Bai,Haizhou Li,Zhiwei Lin,Baosheng Yuan,“Building Class-Based Language Models with Contextual Statistics”,Proc.ICASSP,vol.1,pp.173-176,1998

[4] Peter F.Brown,Vincent J.Della Pietra,Peter V.deSouza,Jenifer C.Lai,Robert L.Mercer,“Class-Based n-gram Models of Natural Language”,Computational Linguistics,Vol.18,No.4,pp.467-479,1992

[5] 小林紀彦, 小林哲則,“自動生成された単語クラスの統計量と単語統計量とを融合した大語彙連続音声認識のための頑強な言語モデル”, 日本音響学会 1999 年春季研究発表会講演論文集.

[6] 寺島志郎, 武田一哉, 板倉文忠,“Bigram 行列の特異値分解による分析”, 日本音響学会 1999 年春季研究発表会講演論文集.