2003－SLP－48　(6)
2003／10／17

HSn-gramの学習法の検討

長野　雄, 鈴木　基之, 伊藤　彰則, 牧野　正三
東北大学大学院工学研究科
〒980-8579 仙台市青葉区荒巻字青葉 05
東北大学 電気・情報系 牧野研究室
e-mail: {nagano,moto,aito,makino}@makino.ecei.tohoku.ac.jp

あらまし　N-gramをHMMに拡張した言語モデルにHSn-gramがある．HSn-gramは，N-gramを決定性有限状態オートマトンとみなし，各状態を複数の状態に分割することで非決定性有限状態オートマトンに拡張したものである．HSn-gramの問題点として，状態数の増加に伴って状態遷移の数が膨大になり，モデルの推定が困難になることがあげられる．本稿ではこの問題点に対処するために，HSn-gram以外のモデルで学習を行い，ある程度パラメータ数を削減した後にHSn-gramの学習を行った．その結果，通常HSn-gramを学習するよりも約5%低いパープレキシティを得ることができた．

キーワード 言語モデル, HS*n*-gram, SS*n*-gram

Examination of the method of learning HSn-gram

Takeshi　NAGANO, Motoyuki　SUZUKI, Akinori　ITO, Shozo　MAKINO
Graduate School of Engineering, Tohoku University
05 Aoba, Aramaki, Aoba-ku, Sendai-shi, 980-8579 Japan
e-mail: {nagano,moto,aito,makino}@makino.ecei.tohoku.ac.jp

Abstract　HSn-gram is a language model which extends an N-gram to Ergodic HMM. HSn-gram regards an N-gram as deterministic finite-state automata, and it extends the FSA into a non-deterministic finite-state automata by dividing each state into two or more states. A problem of learning HSn-gram is that estimation of the model is difficult, because the number of state and the number of state transition becomes large. In this paper, we propose a learning method of an HSn-gram that uses a set of parameters obtained from SSn-gram(the other HMM-based language model) as an initial parameter set. This method reduces the number of parameters, in order to cope with this problem. Consequently, the perplexity is reduced by 5% comparing to that normally learned HSn-gram.

Keywords language model, HS*n*-gram, SS*n*-gram

1 はじめに

N-gram 言語モデルは，音声認識のための言語モデルとして最も一般的に用いられているモデルである．N-gram モデルの中でも，$N = 2$ の bigram と $N = 3$ の trigram モデルが使われることが多い．bigram や trigram は，直前の 1〜2 単語の情報のみを用いて当該単語の確率を推定するため，それ以前の単語の情報を単語生起確率に反映させることができない．N を大きくとればより長い距離の制約が記述可能であるが，確率の推定により大量のサンプルが必要となるため現実的ではない．そのため，N を大きくせずに N-gram に長距離の制約を導入するために可変長 N-gram のアプローチ [1, 2, 3] や長距離にわたる要因によって N-gram を条件づけるアプローチ [4] などの研究がなされてきた．その他のアプローチのひとつに HSn-gram[5] がある．HSn-gram は N-gram を有限状態オートマトンとみなし [6]，これを非決定性有限オートマトンに拡張することによって長距離制約を導入したものである．HSn-gram は状態数を増やすことでより精度のよいモデルを得られることが期待できる．しかし，状態数の増加に伴って状態遷移の数が膨大になり，モデルの推定が困難になるといった問題点がある．本稿ではこの問題点を解決するために HSn-gram の学習法の検討を行った．

2 HSn-gram の概要

一般に，N-gram は決定性有限状態オートマトンとして表現される [6]．例として，語彙を {a, b} としたとき，bigram モデルと等価なオートマトンを図 1 に示す．このとき，各状態は，N-gram における履歴に対応しており，bigram の場合は直前の 1 単語に，trigram の場合は直前の 2 単語に対応する．ここで，各状態を複数の状態に分割し，ある文字列を出力した後の状態が一意に定まらないようにすると，そのオートマトンは非決定性になる．すなわち，このモデルは HMM の一種である．例として，図 1 の各状態を 2 つに分割したものを図 2 に示す．この例では，状態 a が a_1, a_2 に分割され，状態 b は b_1, b_2 に分割されている．現在の状態が a_1 であって，単語 b を出力する場合，状態遷移先は b_1 または b_2 であって，一意に定まらない．このモデルを使って単語列の出現確率を計算する場合，その単語列を生成するすべての状態遷移での遷移確率の総和を計算することになる．

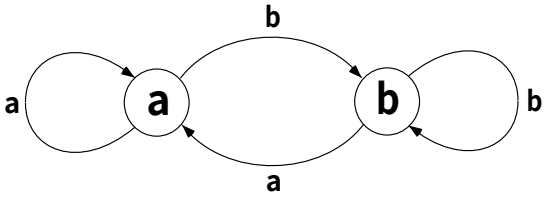

図 1: bigram と等価な有限状態オートマトン

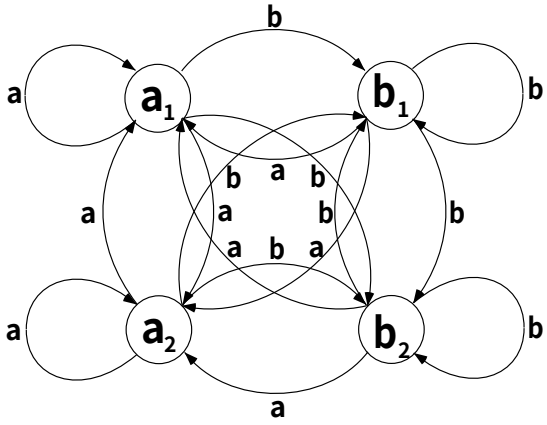

図 2: bigram から生成した非決定性有限状態オートマトン

3 HSn-gram の学習法の検討

HSn-gram の学習は以下の手順で行われる．

1. 非決定性有限状態オートマトンへ変換
   N-gram を有限状態オートマトンに変換する．ただし，一つの履歴に対して複数の状態を割り当てるため非決定性のオートマトンとなる．

2. 状態遷移確率の設定状態遷移確率を N-gram 確率をもとに乱数で設定．

3. モデルの学習 HSn-gram は一種の HMM とみなすことができる．そのため，学習には baum-welch アルゴリズムを用いることが可能である．

学習の手順のうち，1，2 はモデル学習時の「初期モデル」を設定する手順である．HSn-gram は一種の HMM であるため，初期モデルの設定によって性能が異なることが予想される．本稿では学習法の検討として，初期モデルの設定法を検討する．

初期モデルの設定法の検討については，(1) 状態数，状態遷移確率の設定法の検討，(2)HSn-gram 以外の

モデルを HSn-gram の初期モデルに設定，ということが考えられる．

このうち，(1) については文献 [5] で状態数の設定について (i) 履歴によらず状態数を一定で設定，(ii)N-gram 確率をもとにエントロピーで設定，などの方法が検討されている．しかし，どちらの方法で初期モデルを設定しても得られたモデルの性能はほとんど変わらないという結果が得られている．そのため，本稿では HSn-gram 学習のための初期モデルの設定法として，(2) の HSn-gram 以外のモデルで学習を行い，それを初期モデルとして HSn-gram を学習する方法を考える．別なモデルで学習を行うことでパラメータ数を削減しよい初期値を与えることができれば，HSn-gram の学習においてよりよいモデルを得ることが期待される．本稿では HSn-gram の初期モデルを与えるためのモデルとして，SSn-gram[7] を用いることにした．

ここでは，(1)SSn-gram で学習を行う，(2)(1) を初期モデルとして HSn-gram の学習を行う，という手順で HSn-gram の学習を行う．

3.1 SSn-gram の概要

タスクをより小さな複数のサブタスクへ分割し，それぞれのタスクごとに求めた N-gram を適切に切り替えて用いることができれば，性能の高いモデルになると考えられる．SSn-gram は HMM の出力確率を N-gram 確率分布にしたモデルで，学習を行うことで各状態にサブタスクに対応した N-gram を自動的に獲得する．また，その切り替えは HMM の状態遷移確率で表現される．

HSn-gram の初期モデルの設定に SSn-gram を用いることで，次のようなメリットが考えられる．

1. HSn-gram への変換が可能
   SSn-gram は HMM の出力確率を N-gram 化したモデルであり，HSn-gram への変換が可能であると考えられる．

2. より性能のよいモデルが獲得可能
   学習した SSn-gram は通常の N-gram に比べ，よりよい性能を示している [7]．そのため，SSn-gram を初期モデルとすることで，通常の N-gram を種にする文献 [5] の方法に比べ，学習でよりよいモデルが獲得できると考えられる．

次に，SSn-gram から HSn-gram への変換の過程を示す．

3.2 SSn-gram から HSn-gram への変換

SSn-gram から HSn-gram への変換について，次の二つのステップで変換のイメージを説明する．

1. SSn-gram の出力確率 N-gram から HSn-gram への変換

2. 1. から単一の有限状態オートマトンへの変換

ここでは HS-bigram の場合について，以下にステップごとに説明する．

3.2.1 SSn-gram の出力確率 N-gram から HSn-gram への変換

HSn-gram は N-gram を状態遷移図による表現になおし，非決定性のオートマトンに拡張したものであった．まず，最初のステップでは SSn-gram の出力確率に割り当たっている N-gram を有限状態オートマトンへ変換する．この様子を図 3 に示す．図 3 は，出力される記号が x, y の 2 種類だけの場合を示している．図 3 の左側のモデルは SSn-gram を示している．A は状態遷移確率を表しており，A_{11} は状態 1 から状態 1 への遷移確率を示している．左側の SSn-gram の状態 2 から抜き出された右側のモデルは，SSn-gram の出力確率を表している．SSn-gram の出力確率を状態遷移図による表現になおしたものが抜き出された右側のモデルである．抜き出された右側のモデルは N-gram を状態遷移図で表現したものであり，履歴あたりの状態数が 1 である HSn-gram ともいえる．

3.2.2 単一の有限状態オートマトンへの変換

次に，SSn-gram を単一の有限状態オートマトンへ変換する．その様子を図 4 に示す．

図 4 で，『状態 1x』は SSn-gram の状態 1 で出力されていた単語 x を HSn-gram に変換した状態であり，同様に『状態 2y』は状態 2 で出力されていた単語 y を変換した状態を示す．a は HSn-gram の状態遷移確率を表しており，a_{1x1y} は状態 1x から状態 1y への遷移確率を表している．一方，A は SSn-gram に

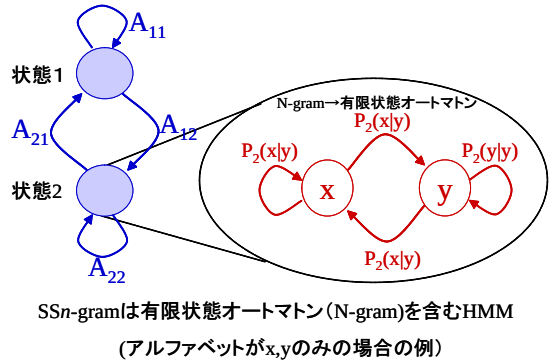

図 3: SSn-gram から HSn-gram への変換–1

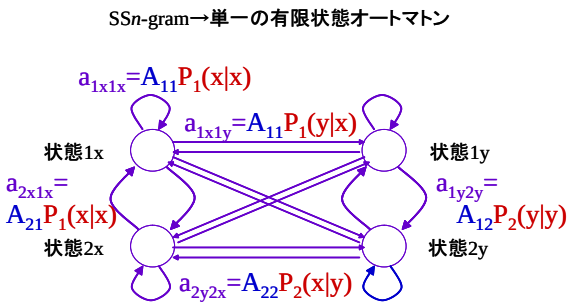

図 4: SSn-gram から HSn-gram への変換–2

おける状態遷移確率を示しており，A_{21} は状態 2 から状態 1 への遷移を表す．また，P は SSn-gram の出力確率を表しており，$P_1(x|x)$ は状態 1 で単語連鎖 (x x) を出力する確率である．

例えば，HSn-gram の状態遷移確率 a_{2x1x} は SSn-gram の状態遷移確率 A と出力確率 P を使って

$$a_{2x1x} = A_{21}P_1(x|x) \tag{1}$$

と表すことができる．

式 (1) の左辺は，HSn-gram において，状態 2x から状態 1x への遷移確率を表している．一方，式 (1) の右辺は SSn-gram において，状態 2 から状態 1 に遷移して単語連鎖 (x x) を状態 1 で出力することを示している．式 (1) のような変換を行うことで，SSn-gram から HSn-gram への変換が行えることを示している．実際には状態からでる遷移確率の和が 1 になるように正規化をしなければならないため，式 (1) で示される HSn-gram の状態遷移確率は厳密ではない．また，アルファベットが二つの場合に限られた式である．

以下に，SSn-gram から HSn-gram への変換の一般式を示す．

$$a'_{iw_kjw_m} = A_{ij}P_j(w_m|w_k) \tag{2}$$

$i = 1 \cdots N_s$，$j = 1 \cdots N_s$，$k = 1 \cdots N_v$，$m = 1 \cdots N_v$ である．

ただし，N_s は SSn-gram の状態数を表し，N_v は語彙数である．また，w_k は語彙中の k 番目の語を表す．a' は正規化前の HSn-gram の状態遷移確率を表している．

式 (2) を使って，正規化された HSn-gram の遷移確率は以下のように表すことができる．

$$a_{iw_kjw_m} = \frac{a'_{iw_kjw_m}}{\sum_{j'=1}^{N_s}\sum_{m'=1}^{N_v} a'_{iw_kj'w_{m'}}} \tag{3}$$

$i = 1 \cdots N_s$，$j = 1 \cdots N_s$，$k = 1 \cdots N_v$，$m = 1 \cdots N_v$ である．

式 (3) によって SSn-gram から HSn-gram への変換が可能である．

4 評価実験

評価実験は HS-bigram で行った．
実験条件を表 1 に示す．

表 1: 実験条件 (small set)

コーパス	毎日新聞
学習テキスト	1994 年 1 月〜9 月 15,000 文
評価テキスト	1994 年 10 月〜12 月 5,000 文
語彙	1994 年 1 月〜9 月 出現頻度上位 5,000 語

実験では初期モデルを与える方法によって得られるモデルの性能の違いを調べた．SSn-gram によって初期モデルを与える方法では，文献 [7] と同様に学習サンプルのクラスタリングを行い，状態数を 2 状態から 5 状態まで変化させて SSn-gram の学習を行った．また，比較のために文献 [5] と同様に履歴によらず一定の状態数を設定し，乱数で初期値を設定した初期モデルとして用いたもので実験を行った．どちらのモデルも以前の報告 [5] と同様に学習，back-off 平滑化を行った．評価はパープレキシティで行った．

表 1 の条件で実験を行った結果を図 5 に示す．図 5 の横軸のパラメータ数は，学習後のモデルの状態遷移の総数である．

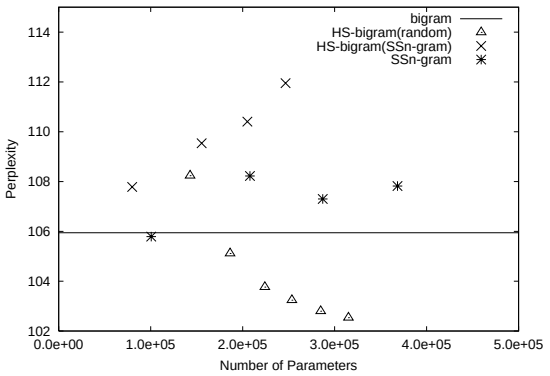

図 5: パラメータ数とパープレキシティ(small set)

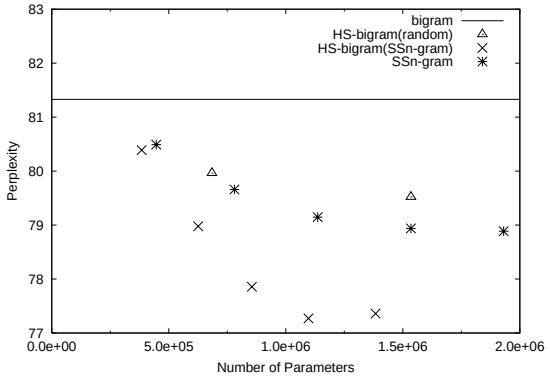

図 6: パラメータ数とパープレキシティ(large set)

図 5 中で,『HS-bigram(random)』は乱数で HS-bigram の初期値を決定したものである．乱数によって初期値を設定したものは HS-bigram の 1 単語あたりの状態数を 2 から 7 まで変化させて実験を行った．また，図 5 中で『HS-bigram(SSn-gram)』は SSn-gram を学習し，それを初期モデルとして HS-bigram を学習したものである．図 5 中の『SSn-gram』は，HS-bigram の初期モデルとした SSn-gram の性能を示している．

図 5 から，初期値を乱数で設定した場合のパープレキシティが低くなった．また，SSn-gram は HSn-gram に比べパープレキシティが高くなった．その結果，SSn-gram を初期モデルとして学習した HSn-gram のパープレキシティが高くなったと考えられる．SSn-gram のパープレキシティが高くなった原因として学習サンプルが十分でないことが考えられる．そこで，表 1 に比べて学習サンプル量を約 5 倍に増やしたデータを使って実験を行った．実験条件を表 2 に示す．

表 2: 実験条件 (large set)

コーパス	毎日新聞
学習テキスト	1994 年 1 月〜9 月約 70,000 文
評価テキスト	1994 年 10 月〜12 月 5,000 文
語彙	1994 年 1 月〜9 月 出現頻度上位 5,000 語

表 2 の条件で実験を行った結果を図 6 に示す．

図 6 中で,『HS-bigram(random)』は乱数で HS-bigram の初期値を決定したものである．乱数によって初期値を設定したものは HS-bigram の状態数が 2 状態の場合と 3 状態の場合について実験を行った．また，図 6 中で『HS-bigram(SSn-gram)』は SSn-gram を学習し，それを初期モデルとして HS-bigram を学習したものである．SSn-gram によって初期モデルを与える方法では，文献 [7] と同様に学習サンプルのクラスタリングを行い，状態数を 2 状態から 6 状態まで変化させて SSn-gram の学習を行った．図 6 中の『SSn-gram』は，HS-bigram(SSn-gram) の初期値とした SSn-gram の性能を示している．

図 6 から，すべてのモデルにおいて bigram よりも性能がよくなった．パラメータ数と性能の関係で比較すると，同じくらいのパラメータ数の場合に HS-bigram(random) に比べて SSn-gram の方が低いパープレキシティが得られた．また，SSn-gram を初期値として学習した HS-bigram(SSn-gram) は，同じパラメータ数の場合に SSn-gram よりもさらに低いパープレキシティが得られた．SSn-gram を学習した結果，通常の bigram に比べてよいパープレキシティが得られており，それを初期モデルとして学習することで HS-bigram(SSn-gram) は結果として低いパープレキシティが得られたと思われる．特にパラメータ数が 1096966 の HSn-gram(SSn-gram) は，bigram に比べて約 5%，初期値とした SSn-gram に比べ約 2%低いパープレキシティを示した．

次に，HS-bigram が長距離制約をどの程度記述できているかを調べるために，評価データの文の長さ別に bigram からのパープレキシティの改善率を調べた．図 6 で一番パープレキシティの低かった HS-bigram(SSn-gram) を使用し，改善率を調べたものを図 7 に示す．

また，この時の評価データの長さ別の出現頻度を図 8 に示す．

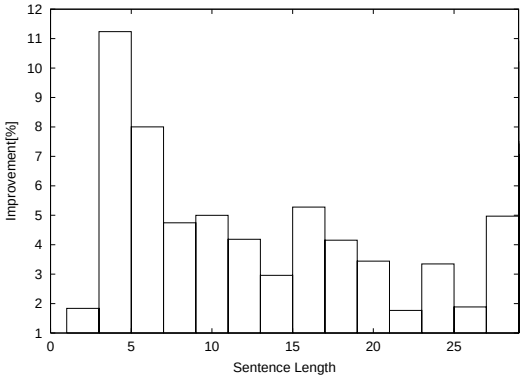

図 7: 文の長さによる改善率の違い

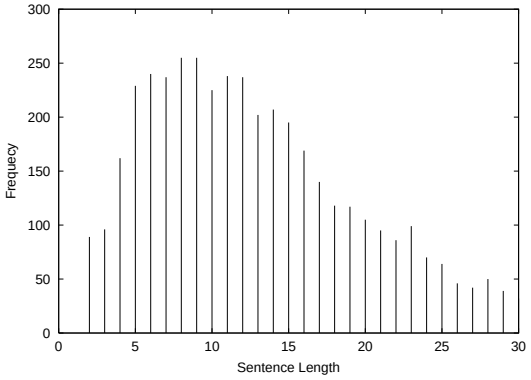

図 8: 文の長さと度数

全評価データでのパープレキシティの改善率は約5%であった．図 7 を見ると，文の長さが長いところでは全体での改善率よりも改善率が低くなっていることがわかる．文の長さが長くなると，長距離制約がない，あるいは長距離制約を記述できていないものが文中で占める割合が多くなることが予想される．そのため，文の長さが長いところでは改善の効果が表れにくくなっていると考えられる．また，図 8 から，文の長さが長いところでは評価データの数が少ないということも考えられる．一方，文の長さが短いところでは改善率が全体での改善率よりも高く，なんらかの長距離制約が記述できていると考えられる．特に長さが 10 以下の文で改善率が高くなった．

5 まとめ

HSn-gram の学習法として，学習された SSn-gram を初期モデルとして HSn-gram を学習する方法を検討した．SSn-gram を初期値とすることで初期値を乱数で決めたものに比べて，同じパラメータ数で比較すると低いパープレキシティが得られた．SSn-gram を初期値とした場合に最もパープレキシティが低くなったモデルでは，bigram に比べて約 5%，初期値とした SSn-gram に比べて約 2%パープレキシティが下がった．また，長さ別に bigram からのパープレキシティーの改善率を見た場合は，文の長さが 10 以下のところで特に改善が大きかった．今後の課題として，trigram による実験や認識実験を行う必要がある．

参考文献

[1] Seymore, K. and Rosenfeld, R.: Scalable backoff language models, *Proc. Int. Conf. on Spoken Language Processing*, Vol. I, pp. 232–235 (1996).

[2] Bonafonte, A. and Mariño, J. B.: Language modeling using x-grams, *Proc. Int. Conf. on Spoken Language Processing*, Vol. I, pp. 394–397 (1996).

[3] Matsunaga, S. and Sagayama, S.: Variable-length language modeling integrating global constraints, *Proc. European Conf. on Speech Comm. and Technology*, pp. 2719–2722 (1997).

[4] Iyer, R. and Ostendorf, M.: Modeling long distance dependence in language: topic mixture vs. dynamic cache models, *Proc. Int. Conf. on Spoken Language Processing*, Vol. I, pp. 236–239 (1996).

[5] 伊藤, 斎藤, 加藤, 好田: N-gram に基づくエルゴディック HMM による言語モデル, 信学技報 SP2000-25, pp. 67–73 (2000).

[6] Hu, J., Turin, W. and Brown, M. K.: Language modeling with stochastic automata, *Proc. Int. Conf. on Spoken Language Processing*, Vol. I, pp. 406–409 (1996).

[7] 長野, 鈴木, 牧野: HMM を用いた複数 n-gram モデルによる言語モデルの構築, 情処論, Vol. 43, No. 7, pp. 2075–2081 (2002).