自然言語処理 118－5
（1997．3．21）

クラス bigram 言語モデルの補間

森 信介
京都大学工学研究科
〒606-01 京都市左京区吉田本町
mori@kuee.kyoto-u.ac.jp

あらまし
本論文では，日本語における単語 bi-gram モデルと品詞 bi-gram モデルの補間を提案する．テストコーパスの解析に必要な未知語モデルも，文字 bi-gram と文字種 bi-gram の補間により得られるモデルで実現する．このモデルの有効性を確かめるため，形態素解析済みのコーパスを用いて単語 bi-gram モデルと品詞 bi-gram モデルとこれらを補間したモデルのテストセットパープレキシティを計算した．その結果，単語 bi-gram モデルでは 151.00 であり，品詞 bi-gram モデルでは 383.61 であり，これらを補間したモデルでは，143.49 であった．単語 bi-gram モデルと品詞 bi-gram を補間したモデルは，単語 bi-gram と同程度の記憶領域で実現できるので，このモデルは単語 bi-gram モデルよりも良いモデルであると結論できる．
キーワード 単語 品詞 n-gram 補間 パープレキシティ

Interpolation of Class-based Bigram Language Models

Shinsuke Mori
Department of Electrical Engineering, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto, 606-01 Japan
mori@kuee.kyoto-u.ac.jp

Abstract
In this paper, we present an interpolated model between a word bi-gram mode and a part-of-speech bi-gram model. We also present, as an unknown word model, an interpolated model between character bi-gram mode and character type bi-gram model. In order to attest an effectiveness experimentally, we calculated perplexities of the word bi-gram model and the part-of-speech bi-gram model and the interpolated model between them. The results of the word bi-gram was 151.00, the part-of-speech bi-gram model, 383.61 and the interpolated model, 143.49. Since the interpolated model needs as large memory space as the word bi-gram model, it follows that the interpolated model is better than the word bi-gram model.
Key Words Word, Part-of-speechl, n-gram, Interpolation, Perplexity

1 はじめに

音声認識などの課題においては，雑音のある通信路を通過することによって，乱れた単語列を復元する課題に直面する．この課題に対処するために，パラメータの推定や実装が容易とされているマルコフモデルを用いることが一般的である．このモデルでは，状態を単語に対応させて，コーパスから状態遷移確率を推定する (単語に基づく n-gram モデル) [1]．しかし，単語を状態に対応させると状態数は語彙数の n 乗と等しくなり，現在一般的に入手可能な量のコーパスでは，状態遷移確率を高い精度で推定できない．その結果，訓練データ以外のデータに対しては，単語列を正確に復元することができない．この問題に対処するため，クラスと呼ばれる単語のグループを一つの状態に対応させることが提案されている [2]．この手法によればモデルの訓練データが十分でない場合にも，性質の似た単語をグループ化することにより状態数が減少し，未観測の単語列に対するモデルの信頼性が向上すると考えられる．このようなモデルを，単語に基づく n-gram モデルに対して，クラスに基づく n-gram モデルと呼ぶ．

クラスに基づく n-gram モデルは単独で用いるだけでなく，単語に基づく n-gram モデルと共に用いることも提案されている [3] [4]．これは，単語に基づく n-gram モデルにおける複数の n-gram モデルの補間を拡張することにより容易に定義できる．すなわち，クラスに基づくモデル n-gram と単語に基づく n-gram モデルを一定の比率で足し合わせるのである．単語に基づく n-gram モデルの補間と同様に，クラスに基づく n-gram モデルは複数あっても良い．また，それぞれの単語をその単語のみからなるクラスとみなすと，単語に基づく n-gram モデルはクラスに基づくモデル n-gram の特殊な場合であるので，クラスに基づくモデル n-gram の補間は単語に基づく n-gram モデルを真に包含することができる．これは，クラスに基づくモデル n-gram と単語に基づく n-gram モデルの補間により得られるモデルが，単語に基づく n-gram モデル以上の能力を持つことを保証する．その一方で，クラスに基づくモデル n-gram と単語に基づく n-gram モデルの補間により得られるモデルは，単語に基づく n-gram モデルよりも多くの状態 (パラメータ) を持つので，その記述にはより大きな領域を必要とする．

本論文では，日本語における単語 bi-gram モデルと品詞 bi-gram モデルの補間を提案し，このモデルの能力をテストセットパープレキシティで評価する．ここで，単語を表記と品詞の直積として定義している．テストセットパープレキシティの計算に必要な未知語モデルは，文字 n-gram で実現することが提案されている [5] が，単語予測の場合と同様に，文字 bi-gram と文字種 bi-gram の補間により得られるモデルで実現した．

上述のモデルの有効性を確かめるため，EDR コーパス [6] を用いて実験を行なった．まず，コーパスを 9 対 1 の比率で学習用とテスト用に分割した．単語 bi-gram モデルと補間クラス bi-gram モデルのパラメータを学習用コーパスで推定し，両方のモデルによるテストコーパスの単語あたりのパープレキシティを計算した．その結果，単語 bi-gram モデルでは 151.00 であり，品詞 bi-gram モデルでは 383.61 であり，これらを補間したモデルでは，143.49 であった．

以下の節では，まず，n-gram 言語モデルについて説明する．次に，実験結果の提示とその評価を行なう．最後に，本研究の結論を述べる．

2 クラス n-gram モデルとその補間

この節では，まず単語 n-gram に基づく確率的言語モデルについて説明する．次いで，その拡張であるクラス n-gram に基づくモデルとそれらの間での補間を説明する．最後に，未知語の扱いについて述べる．

2.1 確率的言語モデル

音声認識などの言語モデルとして，確率的言語モデルがある．これは，対象としている言語の文からその生成確率への関数であり，これは以下の条件を満たす．ただし，Σ は言語のアルファベットであり，M は確率的言語モデルを表わす．

条件 1. $\forall x \in \Sigma^*$ に対して $M(x)$ が定義されている

条件 2. $0 \leq M(x) \leq 1, \quad \forall x \in \Sigma^*$

条件 3. $\sum_{x \in \Sigma^*} M(x) \leq 1$

以上の条件を満たす言語モデルとして，単語 n-gram に基づくモデルがある．これは，文を単語列とみなし，ある時点 i の単語を直前の連続する $n-1$ 個の単語から予測するモデルである．ある文の i 番目の単語を w_i とし，文を $\mathcal{W} = w_1 w_2 \cdots w_m$ とすると，この出現確率は次の式で与えられる．

$$M(\mathcal{W}) = \prod_{i=1}^{m} p(w_i | w_{i-k} w_{i-k+1} \cdots w_{i-1})$$

一般に，確率 $p(w_i | w_{i-k} w_{i-k+1} \cdots w_{i-1})$ の値はコーパスから最尤推定することで得られる．これは，N を単語列

のコーパスにおける頻度として，以下の式で与えられる．

$$p(w_i|w_{i-k}w_{i-k+1}\cdots w_{i-1}) = \frac{N(w_{i-k}w_{i-k+1}\cdots w_i)}{N(w_{i-k}w_{i-k+1}\cdots w_{i-1})} = \frac{N(w_{i-k}w_{i-k+1}\cdots w_i)}{\sum_w N(w_{i-k}w_{i-k+1}\cdots w_{i-1}w)}$$

このように，このモデルは連続する $n=k+1$ 個の単語列の頻度に基づいているので，n-gram モデルと呼ばれる．

確率的言語モデルの評価には，テストコーパスと呼ばれる，パラメータ学習に用いていないある程度の大きさのコーパスの情報量を用いる．これを，クロスエントロピーと呼び，以下の式で定義される．ただし，$C_T=(\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_h)$ はテストコーパスを表わし，$|\mathcal{W}|$ は $\mathcal{W}$ に含まれる文字数を表す．

$$H(C_T, M) = \frac{1}{\sum_{i=1}^{h}|\mathcal{W}_i|}\sum_{i=1}^{h} -\log_2 M(\mathcal{W}_i)$$

音声認識などでは，文字あたりのクロスエントロピーよりもテストセットパープレキシティと呼ばれる値を用いることが多い．これは，一般に単語あたりで計算され，$\overline{|w|}$ を単語の平均文字数として，以下の式で定義される．

$$PP(C_T, M) = 2^{H(C_T,M)\times\overline{|w|}}$$

これは，各単語が等確率に選ばれると仮定した場合の後続可能単語数の幾何平均を表しており，この値が小さいほうがより良い言語モデルである．以下では，テストセットパープレキシティを評価基準とする．

2.2 クラスに基づくモデルへの一般化

単語に基づく n-gram モデルでは，i 番目の単語の予測に直前の長さ $n-1$ の単語列を用いる．このとき，i 番目の単語の確率分布を，長さ $n-1$ の単語列のすべての組合せに対して個々に推定しておき，認識の時に用いる．しかし，これらの直前の単語列のいくつかは，次の単語を予測するという目的においては区別する必要がないという場合がある．このような場合には，直前の事象を一定の長さのすべての単語列に分類することは，不必要に直前の事象を区別していることになる．その結果，限られたコーパスにおける出現回数を減少させ，推定される確率値の信頼性の低下を招く．このような問題を回避する方法の一つとして，あらかじめ単語をクラスと呼ばれるグループに分類しておき，先行するクラスの列を直前の事象とみなして分類するという，クラスに基づく n-gram モデル [2] がある．このモデルでは，次の単語を直接予測するのではなく，次のクラスを予測した上で次の単語を予測する．クラスに基づく n-gram モデルにおける単語列の出現確率は，以下の式で与えられる．

$$p(\mathcal{W}) = \prod_{i=1}^{n} p(c_i|c_{i-k}c_{i-k+1}\cdots c_{i-1})p(w_i|c_i)$$

ここで，c_i は w_i が属するクラスである．単語が複数のクラスに属する場合には，ある単語列を導出するクラス列が複数あることになるので，これらすべてのクラス列に渡って和を計算する必要がある．単語 n-gram に基づくモデルと同様に，確率 $p(c_i|c_{i-k}c_{i-k+1}\cdots c_{i-1})$ の値および，確率 $p(w_i|c_i)$ の値は，コーパスから最尤推定することで得られる．

$$p(c_i|c_{i-k}c_{i-k+1}\cdots c_{i-1}) = \frac{N(c_{i-k}c_{i-k+1}\cdots c_i)}{N(c_{i-k}c_{i-k+1}\cdots c_{i-1})}$$

$$p(w_i|c_i) = \frac{N(w_i, c_i)}{N(c_i)}$$

この式において，単語からクラスへの写像が全単射であれば，単語に基づく n-gram モデルと等価になる．

2.3 低頻度事象への対処

前項で述べたように，n-gram モデルのパラメータ推定には，最尤推定が用いられる．しかし，対象とする事象の頻度が低い場合には，推定値の信頼性は低くなるという問題がある．この問題に対処する方法として，補間と呼ばれる方法が用いられる [7]．これは，次の式で表されるように，より低次のマルコフモデルの遷移確率を一定の割合で足し合わせるという操作を施すことを言う．

- 単語に基づく n-gram モデル

$$p'(w_i|w_{i-k}w_{i-k+1}\cdots w_{i-1}) = \sum_{j=0}^{k}\lambda_j^w p(w_i|w_{i-j}w_{i-j+1}\cdots w_{i-1})$$

ただし $0\le\lambda_j^w\le 1$, $\sum_{j=0}^{k}\lambda_j^w = 1$

- クラスに基づく n-gram モデル

$$p'(c_i|c_{i-k}c_{i-k+1}\cdots c_{i-1}) = \sum_{j=0}^{k}\lambda_j^c p(c_i|c_{i-j}c_{i-j+1}\cdots c_{i-1})$$

ただし $0\le\lambda_j^c\le 1$, $\sum_{j=0}^{k}\lambda_j^c = 1$

係数 λ の値は，確率値 p の推定に用いられるコーパスとは別に用意された比較的小さいコーパスを用いて最尤推定される．この方法では，確率値の推定に用いることができるコーパスの大きさが小さくなり，推定値の信頼性が少しではあるが低下するという問題がある．これに対処する方法として削除補間と呼ばれる方法がある．これは，パラメータ推定のためのコーパスを k 個に分割し，$k-1$ 個の部分で確率値を推定し，残りの部分で補間の係数を推定するということを全ての組合せ (k 通り) に渡って行ない，

その平均値をとるという方法である.

同様の考えに基づいて，複数のクラスに基づく n-gram モデルの間での補間も提案されている [4]. これは，クラスに基づく n-gram モデルが h 個あるとし，それぞれのクラスを $c^1, c^2, \ldots, c^h$ とすると，以下の式で表わされる.

$$p'(w_i|w_{i-k}w_{i-k+1}\cdots w_{i-1}) = \sum_{j=1}^{h} \mu_j p(c_i^j|c_{i-k}^j c_{i-k+1}^j \cdots c_{i-1}^j) p(w_i|c_i^j)$$

ただし $0 \leq \mu_j \leq 1, \sum_{j=1}^{h} \mu_j = 1$

係数 μ_j の値は，確率値 p の推定に用いられるコーパスとは別に用意された比較的小さいコーパスを用いて最尤推定される.

2.3.1 未知語

すでに述べたように，単語 n-gram に基づくモデルは，ある時点の単語を直前の連続する有限個の単語から予測するモデルである. テストコーパスに出現するすべての単語が，学習コーパスに出現することは望めない. 定義から明らかなように，学習コーパスに出現しない単語を含む文の出現確率は 0 となるので，テストセットパープレキシティが無限大となる ($PP = \infty$).

この問題に対処するため，未知語に対応する特別な記号を用意し，未知語はこの記号から生成されることとする. パラメータ推定のための頻度の計数においては，あらかじめ既知語とする単語の集合を決めておき，学習コーパスの既知語以外 (未知語) をこの特別な記号で置き換えてから計数する. また未知語は，文字 n-gram に基づくモデルから生成されることとする. このモデルのパラメータは，特別な記号への置き換えの対象になった単語から推定する. 単語 n-gram モデルの場合と同様に，あらかじめ既知文字とする文字集合を決めておき，それ以外の文字を未知文字を表わす特別な記号で置き換えて文字 n-gram を計数する. 単語 n-gram モデルの場合と異なり，各未知文字は等確率で出現するとする.

クラスに基づく n-gram モデルにおける未知語の対処は，次の 2 つが考えられる.

1. 未知語を表わすひとつのクラスを設ける.
2. 各クラスに対して未知語を表わすクラスを設ける.

形態素解析のなどの応用を考えた場合，品詞をクラスとすることになるが，このときに未知語の品詞 (クラス) を推定する必要がある. 前者の方法では，未知語の品詞が直接推定できないが，後者の方法を採用した場合にはどの品詞の未知語クラスから生成すると確率が最大になるかを探索することで，未知語の品詞が直接推定できるという利点がある. 後者の方法の欠点として，助詞や助動詞などの未知語が非常に少ないクラスの未知語モデルのパラメータの信頼性が低くなることがあげられる. この問題に対処するため，未知語モデルにもクラス (文字種) を導入することが考えられる.

3 実験結果とその評価

我々は，前章で説明した言語モデルを用いて，テストセットパープレキシティを計算した. この章では，この結果を提示し，それに対する評価を述べる.

3.1 実験の条件

実験には EDR コーパス [6] を用いた. まずこれを 10 個に分割し，このうち 9 個を学習コーパスとし，1 個をテストコーパスとした. コーパスに含まれる文数と単語数と文字数を表 1にまとめた. なおアルファベットの数は 6,879 とした. これは，我々の計算機環境で表示可能であった全角文字に文区切り記号を合わせた数である.

表 1: コーパス

用途	文数	単語数	文字数
学習	187,022	4,595,786	7,252,558
評価	20,780	509,261	802,576

実験には，ある時点の記号を直前の記号から予測する bi-gram モデル ($n = 2$) を用いた. 既知語は，2 個以上の学習コーパスに現れる単語 (表記と品詞の直積) とした. 単語 bi-gram に対応するクラスとして，EDR コーパスの 15 の品詞を用いた. 表 2はテストコーパスにおける各品詞の既知語と未知語の数である. 学習コーパスにおける既知語と未知語の数は，コーパスの大きさに比例して，これらの約 9 倍であった. 既知文字は，2 個以上の学習コーパスの未知語集合に現れる文字とした. 文字 bi-gram に対応するクラスとして，我々の直観にしたがって設定した文字種 (数字・平仮名・片仮名・漢字・その他) を用いた. 各 bi-gram モデルのアルファベットは以下の通りである.

- 単語 bi-gram
  1. 既知語 (59,956 個)
  2. 各品詞の未知語を表わす記号 (15 個)
  3. 文区切り記号に対応する状態 (1 個)
- 品詞 bi-gram

表 2: テストコーパスにおける各品詞の既知語と未知語の数

品詞	助詞	名詞	語尾	動詞	記号	助動詞	接尾語
既知語の数	135,634	127,799	60,156	59,670	49,122	30,470	12,671
未知語の数	2	9,048	4	938	13	15	85
未知語率	0.001%	6.612%	0.007%	1.548%	0.026%	0.049%	0.666%

数字	副詞	形容動詞	形容詞	連体詞	接続詞	接頭語	感動詞
7,510	6,868	5,797	5,395	3,773	2,234	2,137	25
358	216	259	84	15	17	25	6
4.550%	3.049%	4.277%	1.533%	0.396%	0.755%	1.156%	19.355%

$$\text{未知語率}(\%) = 100 \times \frac{\text{未知語の数}}{\text{未知語の数+既知語の数}}$$

表 3: 未知語モデルにおける既知文字と既知文字種の数

品詞	助詞	名詞	語尾	動詞	記号	助動詞	接尾語
既知文字の数	7	2,547	8	1,067	62	37	130
既知文字種の数	2	5	3	5	4	3	5

数字	副詞	形容動詞	形容詞	連体詞	接続詞	接頭語	感動詞
46	217	537	172	47	38	39	35
5	5	5	4	3	2	4	4

- 各品詞の既知語を表わす記号 (15 個)
- 各品詞の未知語を表わす記号 (15 個)
- 文区切り記号に対応する状態 (1 個)

• 各品詞の文字 bi-gram

1. 既知文字 (表 3)
2. それ以外の文字種 (1 個)
3. 単語区切り記号に対応する状態 (1 個)

• 各品詞の文字種 bi-gram

1. 既知文字種 (表 3)
2. それ以外の文字種 (1 個)
3. 単語区切り記号に対応する状態 (1 個)

補間係数の推定には削除補間法を用いた．すなわち，9 個の学習コーパスのうちの 8 個で記号列の頻度を計数し，残りの 1 個の出現確率が最大になる補間係数の推定を 9 通り行ない，その平均値を補間係数とした．この補間係数と全ての学習コーパスに対して計数した状態列の頻度をパラメータとするマルコフモデルを構成し，テストセットパープレキシティを計算した．テストコーパスの単語区切りや品詞は，コーパスにあらかじめ付加されたものを用いた．したがって，テストコーパスに含まれる文字列の出現確率は，その文字列のすべての生成方法による確率を合計した値ではなく，コーパスに示された生成方法のみによる値である．なお，単語区切りが明示されていない文に対しては，動的計画法を用いたアルゴリズムにより，出現確率が最大となる単語区切りを，文に含まれる文字数に比例した時間で求めることができる [8].

3.2 結果と考察

単語 bi-gram モデルと品詞 bi-gram モデルとそれらを補間したモデルの能力を調べるために，それぞれのモデルのテストセットパープレキシティの計算を行なった．単語 bi-gram モデルの未知語モデルは文字 bi-gram モデルであり，品詞 bi-gram モデルの未知語モデルは文字種 bi-gram モデルであり，単語 bi-gram モデルと品詞 bi-gram モデルを補間したモデルの未知語モデルは，文字 bi-gram モデルと文字種 bi-gram モデルを補間したモデルである．表 4は，それぞれのモデルのテストセットパープレキシティである．この結果から，単語 bi-gram モデルは品詞 bi-gram モデルよりもかなりよいモデルであることが分かる．ただし，単語 bi-gram モデルは品詞 bi-gram モデルに比べて非常に大きな記憶容量を要することに注意しなければならない．これらを補間したモデルのパープレキシ

表 4: 各モデルの単語あたりのテストセットパープレキシティ

モデル	パープレキシティ
単語 bi-gram モデル	151.00
品詞 bi-gram モデル	383.61
単語 bi-gram と品詞 bi-gram の補間モデル	143.49

ティは，単語 bi-gram モデルのパープレキシティよりも低くなっている．必要となる記憶容量の大きさはあまり変わらないので，このモデルは単語 bi-gram モデルよりも良いモデルであると結論できる．

4 おわりに

本論文では，日本語における単語 bi-gram モデルと品詞 bi-gram モデルの補間を提案し，このモデルの能力を単語 bi-gram モデルや品詞 bi-gram モデルと実験的に比較した．評価には，テストセットパープレキシティを用いた．その結果，単語 bi-gram モデルでは 151.00 であり，品詞 bi-gram モデルでは 383.61 であり，これらを補間したモデルでは，143.49 であった．単語 bi-gram モデルと品詞 bi-gram を補間したモデルは，単語 bi-gram と同程度の記憶領域で実現できるので，このモデルは単語 bi-gram モデルよりも良いモデルであると結論できる．

参考文献

[1] C. E. Shannon. Prediction and Entropy of Printed English. *Bell System Technical Journal*, Vol. 30, pp. 50–64, 1951.

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

[3] F. Jelinek. Self-Organized Language Modeling for Speech Recognition. Technical report, IBM T. J. Watson Research Center, 1985.

[4] John G. McMahon and Francis J. Smith. Improving Statistical Language Model Performance with Automatically Generated Word Hierarchies. *Computational Linguistics*, Vol. 22, No. 2, pp. 217–247, 1996.

[5] 永田昌明. EDR コーパスを用いた確率的日本語形態素解析. EDR 電子化辞書利用シンポジウム, pp. 49–56, 1995.

[6] 日本電子化辞書研究所. EDR 電子化辞書仕様説明書, 1993.

[7] Fredelick Jelinek, Robert L. Mercer, and Salim Roukos. Principles of Lexical Language Modeling for Speech Recognition. In *Advances in Speech Signal Processing*, chapter 21, pp. 651–699. Dekker, 1991.

[8] H. Ney. The Use of One-Stage Dynamic Programming Algorithm for Connected Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 32, No. 2, pp. 263–271, 1984.