音声言語情報処理 19－13
(1997. 12. 12)

再現性を考慮した文字列に基づく統計的言語モデル

○森 大毅† 阿曽 弘具† 牧野 正三‡

† 東北大学大学院工学研究科
‡ 東北大学大型計算機センター

〒 980-77 仙台市青葉区荒巻字青葉

あらまし 本報告では，知識に依存しない，高い曖昧性削減能力を持つ新しい言語モデルを提案する．このモデルはsuperwordと呼ぶ文字列の集合の上のn-gramとして定義され，従来の単語や文字列のn-gramモデルを包含するものになっている．superwordは訓練テキスト中の文字列の再現性のみに基づいて定義される概念であり，Forward-Backwardアルゴリズムによって学習される．実験の結果，superwordに基づくモデルと文字のtrigramモデルを複数融合させたモデルの優位性が示され，形態素解析に基づく方法を上回る性能が得られた．

キーワード 言語モデル，superword，n-gram，音声認識，文字認識

Natural Language Models Based on Repetitional String

Hiroki Mori†, Hirotomo Aso†, and Shozo Makino‡

†Graduate School of Engineering, Tohoku University
‡Computer Center, Tohoku University
Aoba, Aramaki, Aoba-ku, Sendai-shi, 980-77 Japan

Abstract In this report, a new, knowledge-free language model with great ability in reducing ambiguity. This model is defined as n-gram of string which is referred to "superword," and belongs to a superclass of traditional word or string n-gram models' class. The concept of superword is based on only one principle—repetitionality in training text. The probabilistic distribution of the model is learned through the forward-backward algorithm. Experimental results showed that the performance of superword model combined with character trigram model was superior to the traditional word model based on morphological analysis.

key words language model, superword, n-gram, speech recognition, character recognition

1. はじめに

音声認識・文字認識の精度向上のため，より高い性能を持つ言語モデルを求めることは重要である．近年は，モデル構築やメンテナンスの容易さの点から，コーパスに基づく統計的言語モデルの研究が盛んである．大語彙ないしタスク非依存のシステムのための統計的言語モデルとして今日もっとも有望視されているものに，n-gram が挙げられる．n-gram は大量のテキストコーパスからの単純な数え上げによって得られる統計量であり，強力かつ頑健性に優れている．

英語などのヨーロッパ系言語においては，n-gram の単位として，単語あるいは品詞に代表される単語クラスを用いることが多い．これらの言語においては単語は分かち書きされるため機械的に取り出すことができ，数え上げも容易に行える．

これに対し，日本語や中国語には分かち書きの習慣がない．朝鮮語は文節ごとに分かち書きをするが，その分かち方は一定しないうえ，n-gram の単位としては大き過ぎて汎化性に難がある．よって，これらの言語を n-gram によってモデル化する際には，テキストコーパスになんらかの前処理が必要である．これには次の可能性が考えられる．

- 人手によって解析されたコーパスを使う
- 自動形態素解析システムによって単語に分割する
- 経験的な統計基準によって文字列に分割する

このうち解析済みコーパスを使う方法には，コーパス自体の入手が質的・量的な困難を伴うという欠点がある．形態素解析に基づく方法は有効であるが，モデルを学習するためにはまず形態素解析システムを用意せねばならないうえ，特定タスクに対して高い性能を得るためには予め辞書をチューニングする必要があると考えられ，メンテナンスのコストがかかる．また，形態素解析システムの文法規則によっては機能語が短めに分割される傾向があり，n-gram の性能を必ずしも最大にするものではない．

これらの手法に対して，伊藤ら[1]は統計的な基準によって文字列の集合を選定し，その文字列に分割されたテキストを使って n-gram を学習する方法を提案している．文字列を選定する基準としては，単純な頻度，および語彙の自動獲得のために提案されている正規化頻度[2]の高いものから選ぶ方式が有効であったとされる．この方法は，形態素解析を必要としない点で優れている．しかし，抽出すべき文字列の最適な個数を見出す方法については述べられていない．また，用いられている基準と言語モデルの能力との理論的関係は浅く，最良の分割方法である保証はない．さらに，この手法ではテキストが明示的に分割される．このため，接辞を伴った語や複合語などの長い文字列が抽出された場合，その文字列を構成するもっと短い語は出現しなかったのと同様な扱いを受けることになる．有限のテキストから汎化性の高い言語モデルを構築したい場合に，このような明示的な分割が最良の結果を与えるとは限らない．

本報告では，高い曖昧性削減能力を持つ新しい言語モデルを提案する．このモデルは，superword と呼ぶ文字列の集合の上の n-gram モデルとして定義される．superword は訓練テキスト中の文字列の再現性のみに基づいて定義される概念であり，与えられた訓練テキストに対して一意に定まる．具体的な確率分布は，訓練テキストから Forward-Backward アルゴリズムによって求める．訓練テキストを明示的に分割せぬまま学習を行うため，長い文字列中の部分文字列を「再利用」することが可能となり，少量の訓練テキストでも効率の良いモデル化が期待できる．本報告ではまた，いくつかのモデルの融合による汎化性の向上についても検討する．

2. superword モデルの定式化

単語[3]や文字列の n-gram[1]では与えられた系列を単語ないし文字列に分割するやり方が一意に決まらないため，これらのモデルは直前の $n-1$ 個の単語や文字列を状態とする，隠れマルコフモデルの一種と考えられる．単語や文字列の集合は，語彙知識として人手で与えられるか，あるいは経験的な規則に基づいて訓練テキストから抽出されるものである．ここで定義する superword とはこれら単語や文字列を一般化したものであるが，それらと対照的なのは，訓練テキスト中の任意の文字列を含み得る点である．ただし，言語モデルとして意味を持つために必要最小限のヒューリスティクスは必要である．そこで，次の条件を満たす文字列を superword と定義する．

- 訓練テキスト中に最低 2 回出現する

または

- 長さ 1 の文字列である

訓練テキストにおける再現性の仮定は，ある文字列が何らかの言語的なまとまりを成すか否かに対する基準となるものであり，そのような基準として考

え得る制約の中でもっとも緩い条件として与えてある．また，再現性とは独立に，長さ1の文字列は全てsuperwordと定義している．これにより，全ての文は少なくとも1通りのsuperwordの系列として表現できることが保証される．superword n-gram確率 $P(w_i|w_{i-(n-1)}\cdots w_{i-1})$ は，直前に $n-1$ 個のsuperwordの列 $w_{i-(n-1)}\cdots w_{i-1}$ が生起したと仮定した時のsuperword w_i の条件付き生起確率である．

与えられた文 $\boldsymbol{C}=C_1C_2\cdots C_k$ がsuperwordの列 $w_1w_2\cdots w_l$ に分割できるとき，$w_1w_2\cdots w_l \in \boldsymbol{C}$ と書く．superword n-gramモデルは，$\boldsymbol{C}$ の全ての可能な分割に関して計算したsuperword n-gram確率の積の総和をもって $\boldsymbol{C}$ の発生確率を推定するものである．すなわち，その確率を次式で与える．

$$P(\boldsymbol{C})=\sum_{w_1\cdots w_l\in\boldsymbol{C}}\prod_{i=1}^{l}P(w_i|w_{i-(n-1)}\cdots w_{i-1}) \quad (1)$$

ここで $n=1$ の時，すなわちsuperword unigramモデルは，文全体の生起確率がそれぞれ独立なsuperwordの生起確率の積で表されるとするものであり，multigram[4]と呼ばれる可変長単語列に基づく言語モデルと同一のものである．また，superword n-gramモデルのクラスは，単語や文字列の n-gramモデルのクラスを包含する．

3. superwordモデルの学習法

3.1 superword集合の獲得

モデルの獲得にあたっては，パラメータの学習に先立ち，訓練テキストからsuperwordの集合を求める必要がある．長さ1のsuperwordについては自明であるから，再現性のある文字列を集める作業が核心である．これには，訓練テキストの全ての位置から始まる半無限文字列をソートして任意長 n-gram統計を求め[5]，2回以上出現する文字列を記録する方法が考えられる．しかし，再現性のある文字列だけに興味がある場合には，短い文字列から長い文字列へと逐次的に求める簡便な方法で十分である[6]．

実験で用いたテキストコーパスでは，長さ L のsuperwordの種類は大きな L では単調に減少することが観察されている．

3.2 確率分布のForward-Backward学習

superwordモデルでは，ある状態から別の状態に移る時に，ある確率で一つのsuperwordを出力する．w_i の表記を $C_1\cdots C_j\cdots C_L$ とし，w_i の長さ j のプレフィックスを $w_{i,j}$ とする．図1に示すように，確率 $P(w_i|w_{i-(n-1)}\cdots w_{i-1})$ で副状態

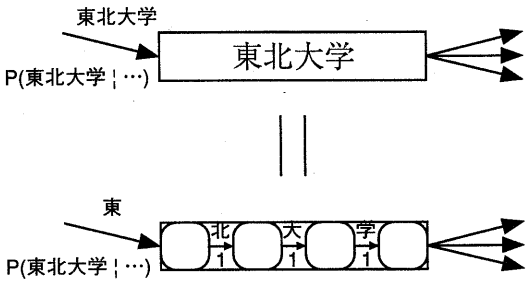

図1. 「東北大学」というsuperwordの各文字に対応した副状態の系列

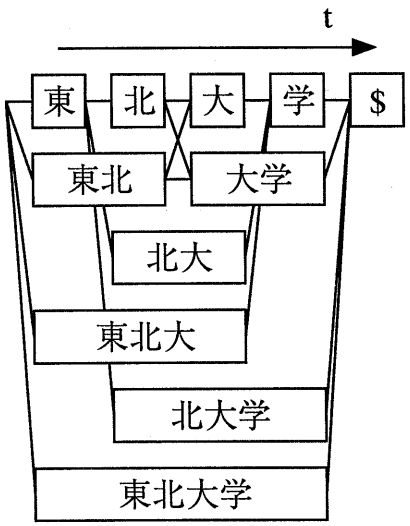

図2. 「東北大学」というテキストの解析．矩形はsuperwordを，実線は可能なパスを表す．“$”は文の終端

$w_{i-(n-2)}\cdots w_{i-1}w_{i,1}$ に移る時に C_1 を出力し，以後確率1で副状態 $w_{i-(n-2)}\cdots w_{i-1}w_{i,j}$ に移る時に C_j を出力し，最終的に状態 $w_{i-(n-2)}\cdots w_{i-1}w_i$ に至ると考える．すなわち，単位時間で1回の状態遷移をし，記号を1つ出力する通常の隠れマルコフモデルと同様に扱うことができる．

$n=1$，すなわちsuperword unigram確率の学習のための初期確率としては，全てのsuperwordが等確率で発生するとして，superwordの数の逆数を与える．$n>1$ については，対応するsuperwordの $(n-1)$-gram確率で初期化する．

確率の再推定のために，図2のように訓練テキストから全てのsuperwordを洗い出す．次に，連接可能な n 個のsuperwordの組に関して，次式によって確率を更新する．

$$\tilde{P}(w_i|w_{i-(n-1)}\cdots w_{i-1}) = \frac{\sum_t \alpha_{t-1}(w_{i-(n-1)}\cdots w_{i-1})P(w_i|w_{i-(n-1)}\cdots w_{i-1})\beta_t(w_{i-(n-2)}\cdots w_i)}{\sum_t \alpha_t(w_{i-(n-2)}\cdots w_i)\beta_t(w_{i-(n-2)}\cdots w_i)} \quad (2)$$

ただし，α, βはそれぞれ Forward 確率，Backward 確率で，以下のように再帰的に定義する．

$$\alpha_1(w) = P(w|\#), \quad \#は文頭を表す状態 \quad (3)$$

時刻 $t(t>1)$ で superword w_iの第 1 字目を出力するとき

$$\alpha_t(w_{i-(n-2)}\cdots w_i) = \sum_{w_{i-(n-1)}} \alpha_{t-1}(w_{i-(n-1)}\cdots w_{i-1})P(w_i|w_{i-(n-1)}\cdots w_{i-1}) \quad (4)$$

時刻 $t(t>1)$ で superword $w_i = C_1\cdots C_j\cdots C_L$の第 j字目 $(j>1)$ を出力するとき

$$\alpha_t(w_{i-(n-2)}\cdots w_{i-1}w_{i,j}) = \alpha_{t-1}(w_{i-(n-2)}\cdots w_{i-1}w_{i,j-1}) \quad (5)$$

ただし

$$\alpha_t(w_{i-(n-2)}\cdots w_{i-1}w_i) = \alpha_t(w_{i-(n-2)}\cdots w_{i-1}w_{i,L}) \quad (6)$$

同様に

$$\beta_T(\$) = 1, \quad T は文末記号 \text{“\$”} を出力する時刻 \quad (7)$$

時刻 $t(t<T)$ で superword w_iの第 1 字目を出力するとき

$$\beta_{t-1}(w_{i-(n-1)}\cdots w_{i-1}) = \sum_{w_i} \beta_t(w_{i-(n-2)}\cdots w_i)P(w_i|w_{i-(n-1)}\cdots w_{i-1}) \quad (8)$$

時刻 $t(t<T)$ で superword $w_i = C_1\cdots C_j\cdots C_L$の第 j字目 $(j>1)$ を出力するとき

$$\beta_{t-1}(w_{i-(n-2)}\cdots w_{i-1}w_{i,j-1}) = \beta_t(w_{i-(n-2)}\cdots w_{i-1}w_{i,j}) \quad (9)$$

ただし

$$\beta_t(w_{i-(n-2)}\cdots w_{i-1}w_i) = \beta_t(w_{i-(n-2)}\cdots w_{i-1}w_{i,L}) \quad (10)$$

4. 長さ制限の導入

再現性のある文字列の長さを十分大きく取れば，前節までに述べたモデルは与えられた訓練テキストに対して一意に求まる．以下では，これを一般 superword n-gram モデルと呼ぶ．しかし，一般モデルのパラメータ数は大きい．特に，$n>2$ では superword の組み合わせが爆発し，現実的ではない．さらに，あまりに長い superword は訓練テキストに特化してしまう恐れがあり，汎化能力の低下を招く．

これに対処するため，一般モデルに加えて長さ制限付きの superword モデルを導入する．これは，逐次的な再現性文字列の獲得を早い段階で打ち切って小さな superword の集合をつくり，その集合に基づいて Forward-Backward 学習を行うことで得ることができる．以下では，長さ lに制限された superword n-gram 確率を $P_{|w|\leq l}(w_i|w_{i-(n-1)}\cdots w_{i-1})$ と表記する．

5. 複合モデル

n-gram に代表される確率モデルにおいては，モデルのパラメータを精度良く推定するに足るサンプルが得られないことが多く，パラメータ空間のさまざまなスムージング法が提案されている[7]．その一つに，いくつかのモデルの確率の重み付き線形和で表現する方法がある[8]．これは本来，詳細なモデルの値が信用できない場合に，パラメータの少ない安定したモデルの値を代用するものであるが，性質の異なる複数のモデルを組み合わせてより良いモデルを得るという積極的な利用も可能である．本節ではいくつかの複合モデルを考える．

superword bigram$(n=2)$ モデルに対しては，superword unigram 確率によって補間された確率は次式で与えられる．

$$\hat{P}(w_i|w_{i-1}) = \lambda_g P(w_i|w_{i-1}) + (1-\lambda_g)P(w_i) \quad (11)$$

重み係数λ_gは，訓練テキストとは別のサンプル (held-out データ) またはクロスバリデーションによって得られる仮想的な未知データの確率を最大にするように再推定する．

前述したように，一般 superword bigram はパラメータ量が多くなり過ぎるので，実際には superword の長さを最大 lに制限したモデルと組み合わせる．これは次式で与えられる．

$$\hat{P}_{|w|\leq l}(w_i|w_{i-1}) = \lambda_b P_{|w|\leq l}(w_i|w_{i-1}) + (1-\lambda_b)P_{|w|\leq l}(w_i) \quad (12)$$

式(12)のような制限されたモデルでは，長い語の表現に難があることも考えられる．そこで，長さ制限付き superword bigram モデルと一般 superword unigram モデルの複合モデルを導入する．複合 superword bigram 確率は次式で定義される．

$$P_{\mathrm{comp}}(w_i|w_{i-1}) = \lambda_{\mathrm{c}}\hat{P}_{|w|\leq l}(w_i|w_{i-1}) + (1-\lambda_{\mathrm{c}})P(w_i) \quad (13)$$

さらに，複合 superword bigram モデルを，文字の trigram モデルによってスムージングすることを考える．文字の trigram モデルは，それ自身で強力な曖昧性削減能力を持っているが[9]，単語 n-gram モデルと融合させることにより，認識対象中の未知の文字列の存在による解析精度の低下の影響を低減させ，頑健なモデルとすることができる[3]．文字によって補間された複合 superword bigram 確率は次式で定義される．

$$\hat{P}_{\mathrm{comp}}(w_i|w_{i-1}) = \lambda_{\mathrm{w}}P_{\mathrm{comp}}(w_i|w_{i-1}) + (1-\lambda_{\mathrm{w}})\hat{P}_{\mathrm{c}}(w_i|w_{i-1}) \quad (14)$$

ただし，$\hat{P}_{\mathrm{c}}(w_i|w_{i-1})$ は superword w_iが生起する確率を，補間された文字 trigram 確率の積によって求めたものである．すなわち，w_iの表記を $C_1\cdots C_{L(w_i)}$，w_{i-1}の最後の2文字を $C_{-1}C_0$ と書くとき

$$\hat{P}_{\mathrm{c}}(w_i|w_{i-1}) = \left(\prod_{j=1}^{L(w_i)}\hat{P}_{\mathrm{c}}(C_j|C_{j-2}C_{j-1})\right)\cdot d(L(w_i)) \quad (15)$$

ただし$\hat{P}_{\mathrm{c}}(C_j|C_{j-2}C_{j-1})$ は bigram, unigram 等により補間された文字 trigram 確率である．また，$d(L(w_i))$ は文字モデルが生成する単語の長さに関する分布関数である．

6. 評価実験

提案した言語モデルの能力を，文字を単位としたパープレキシティによって評価する．パープレキシティは，式(1)において評価用テキストを$\boldsymbol{C}$として次式で求められる．

$$PP \simeq \hat{P}(\boldsymbol{C})^{-1/k} \quad (16)$$

ただし，kは評価用テキストの全字数である．長さ1の superword に対しては，確率が設定した底値を下回る場合には底上げした．対象タスクは朝日新聞「社説」とした．実験に用いたテキストの量を表1に示す．表中，held-out とは式(12), 式(13), 式(14)の重み係数を求めるために用いたテキストである．各々のテキストは，共通部分を持たない．

長さ制限の効果を見るため，superword unigram モデルについて最大長を変化させてパープレキシティを求めた．その結果を図3に示す．この結果から，長い superword を許してもパープレキシティは上がらないことがわかる．これは，superword の再現性の条件が適当であったことを示す．以下の実験では，$L=20$ の場合を一般 superword unigram モデルとして扱う．

表1. 訓練テキスト・評価テキストの量

	字	(文)
訓練	969497	(21767)
held-out	85654	(1953)
評価	80098	(1779)

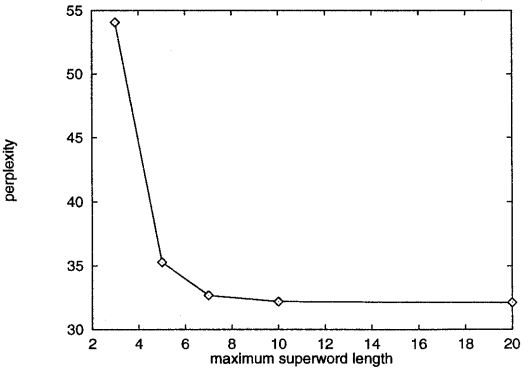

図 3. superword unigram モデルにおける長さ制限の効果

表2に，提案したモデルのパープレキシティを示す．上から4項目までが superword に基づくモデルである．bigram とあるのは式(12)の長さ制限付きモデルである．ここでは最大長を3とした．一般 unigram+bigram とあるのは式(13)の複合モデル，一般 unigram+bigram+文字とあるのは式(14)のさらに文字 trigram で補間したモデルである．その場合の式(15)の分布関数としては，指数分布を仮定した．表2の残りの3項目は比較のために示してある．単語 trigram は，訓練テキストをあらかじめ形態素解析システム JUMAN[10]により分割して求めたものであり，削除補間法によりスムージングしたものである．文字+単語 trigram は，さらに文字の trigram でスムージングしたもので，式(15)と同様の式を用いている．

この結果から，次のことがわかる．まず，superword unigram モデルの性能が良くない．図3の結果をも考慮すると，これは superword の長さの問題ではなく，unigram では語と語の連接関係が本質的に表現できないものと考えられる．これは ATIS デー

表 2. 各モデルの性能評価

モデル	パープレキシティ
unigram	32.4
bigram	29.8
一般 unigram+bigram	28.5
一般 unigram+bigram+文字	25.7
文字 trigram	28.9
単語 trigram	28.6
文字+単語 trigram	26.6

タベースの上での multigram の評価[4]といくぶん矛盾する結果であるが，伊藤ら[1]も同様の結果を導いている．

長さ制限 superword bigram モデルの導入によって，性能の向上が見られた．しかし，まだその性能は文字 trigram モデルに及ばない．

長さ制限 superword bigram モデルと一般 superword unigram モデルを融合させることで，若干の性能向上が見られた．これは，長い superword は単独ではあまり性能に貢献しないが，語と語の連接関係だけでは表現しきれない部分を補う効果を持っているものと考えることができる．語と語の関係に関する知識と語彙知識とを独立に表現する枠組は，形態素解析の原理と類似している．

さらに，文字 trigram モデルでスムージングすることにより，大きく性能が向上した．その結果，形態素解析を用いたモデルを超える性能が得られた．superword に基づいたモデル単独では訓練テキストに対して過学習する傾向があり，未知テキストに対して脆弱な側面があるが，未知テキストに対して頑健な文字 trigram モデルとの融合によりそれが克服できることを意味する．

7. あとがき

本報告では superword の概念に基づいた言語モデルを獲得する方法について述べた．評価実験の結果，長さ制限を施した superword bigram モデルを文字 trigram モデルと組み合わせて頑健性を向上させたモデルの性能が高く，形態素解析に基づく手法を超える能力が得られた．

改善すべき点として，訓練テキストに比べモデルの規模が大きいことが挙げられる．学習により小さな確率を付与された状態遷移の枝刈りなどにより，性能を維持したままコンパクトなモデルとする必要があろう．

参考文献

[1] 伊藤彰則，好田正紀，“かな・漢字文字列の連鎖統計による言語モデル，” 信学論 (D-II), vol.J79-D-II, no.12, pp.2062–2069, 1996.

[2] 中渡瀬秀一，“統計的手法による単語の切出しについて”，信学技報，NLC95-68，1995.

[3] 森　大毅，阿曽弘具，牧野正三，“文字・単語 n-gram の融合に基づく言語モデル，” 信学技報，NLC96-24，1996.

[4] S. Deligne and F. Bimbot, “Language modeling by variable length sequences: Theoretical formulation and evaluation of multigrams,” Proc. ICASSP 95, pp. 169–172, 1995.

[5] M. Nagao and S. Mori, “A New Method of N-gram Statistics for Large Number of N and Automatic Extraction of Words and Phrases from Large Text Data of Japanese,” Proc. 15th International Conference on Computational Linguistics, pp.611–615, 1994.

[6] 森　大毅，阿曽弘具，牧野正三，“再現性 n-gram 統計の効率的な構成法，” 1996 信学ソ大，D-56，1996.

[7] M. Federico, M. Cettolo, F. Brugnara and G. Antoniol, “Language modelling for efficient beam-search,” Computer Speech and Language, vol.9, pp.353–379, 1995.

[8] F. Jelinek and R.L. Mercer, “Interpolated estimation of Markov source parameters from sparse data,” in Pattern Recognition in Practice, eds. E.S. Gelsema and L.N. Kanal, pp.381–397, North-Holland, Amsterdam, 1980.

[9] H. Mori, H. Aso, and S. Makino, “Robust n-Gram Model of Japanese Character and Its Application to Document Recognition,” IEICE Trans. on Information and Systems, vol.E79-D, no.5, pp.471–476, 1996.

[10] 松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木　裕, 長尾　真, “日本語形態素解析システム JUMAN 使用説明書 2.0,” 奈良先端大技報, NAIST-IS-TR94025, 1994.