自然言語処理 114－16
(1996. 7. 19)

文字・単語 n-gram の融合に基づく言語モデル

○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 べた書きのテキストコーパスから構築することのできる言語モデルとしては、文字を単位とした n-gram モデルが有効であることが知られている。しかし、さらに強い制約を得るために n-gram の単位を文字から単語に拡張すると、単語境界の曖昧性や少数サンプルの影響が無視できない。本報告では、削除補間法に基づいて単語 n-gram モデルを文字 n-gram モデルと融合させることにより、これらの問題を解決した言語モデルを提案する。パープレキシティを基準とした評価実験により、提案するモデルが他のモデルに比べ高い曖昧性削減能力を持つことを示す。

キーワード 言語モデル、n-gram, 削除補間法

Natural Language Models Based on Combination of Character and Word n-grams

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 n-gram model of character is known to be an effective language model that can be obtained from plain text corpora. To have more effective model, extension from character to word n-gram will be desired. As for Japanese, however, the extension cannot be straightforward because of ambiguity of word segmentation, and the problem of sparse data. In this report, we propose a new language model based on the combination of character and word n-gram model with *deleted interpolation method*. The proposed model's superiority in reducing ambiguity is revealed through several experiments.

key words language model, n-gram, deleted interpolation method

1. はじめに

文字認識・音声認識などにおける曖昧性の削減のために、言語モデルの利用が有効であることが知られている。特に近年は、コーパスに基づく言語モデルがさかんに用いられるようになった。これは計算機の性能向上と、言語資源の充実に負うところが大きい。しかし、EDR コーパスのようなタグ付きのコーパスの整備はまだ広い分野で使えるようになってはいない。また、日本語を記述するための形態素や文法として利用できる体系には標準と呼べるものは存在しない。このため、異なった文法体系に基づいたコーパスを同時に用いることが難しい。これに比べ、タグのないべた書きのテキストは、文章を計算機上で処理する場合の基本的な形式であるために、さまざまな分野のものがコーパスとして利用できることが期待される。したがって、言語モデルもべた書きのテキストから自動獲得できることが望ましい。

文字を単位とする n-gram モデル (マルコフモデル) は、べた書きテキストから構築することができる点で優れている。日本語の書き言葉においては文字 n-gram は比較的強い制約を与え、文字認識の曖昧性解消にも有効であることがわかっている[1]。しかし、n-gram モデルでは n の指数オーダーのパラメータを推定せねばならないため、n を大きくすることで制約を強めることが困難である。

テキストコーパスを用いてより強い制約を持つ言語モデルを得たいと考えた場合に、n-gram の単位を文字より大きい単位、例えば単語単位に拡張したいと考えるのは自然であろう。しかし、日本語の場合は単語の境界を明確に定めることができないために、文字 n-gram の場合にはなかった新たな曖昧性が生まれる。これを以下では単語境界の曖昧性と呼ぶ。テキストコーパスから単語 n-gram を学習する段階、およびそれを用いて認識を行う段階の両方において、曖昧性は必然的に付随する。さらに、単語を単位とすることでカテゴリ数が文字の場合よりも大きくなるため、サンプルが相対的に少数になってしまい、適切なモデル構築への影響が顕著になることが考えられる。

本報告では、単語 n-gram モデルにおけるこのような問題点を、文字 n-gram モデルとの融合によって解決できることを示す。融合の手法としては、少数サンプルの場合にスムージングの手段として用いられる削除補間法[2]を拡張して適用することを提案する。また、この融合モデルが高い曖昧性削減能力を持つことを、他の手法とのパープレキシティの比較によって示す。

2. 言語モデルの定式化

ここでいう言語モデルとは、文字の系列 $\boldsymbol{C} = C_1C_2\cdots C_k$ に対してその生起確率 $P(\boldsymbol{C})$ の推定値 $\hat{P}(\boldsymbol{C})$ を与えるものである。

2.1 文字 n-gram モデル

言語モデルとして、まず文字 n-gram モデルを説明する。文字 n-gram 確率とは、直前 $n-1$ 文字のコンテキストに関する、文字の条件付き生起確率 $P(C_i|C_{i-(n-1)}\cdots C_{i-2}C_{i-1})$ である。文字 n-gram モデルは、$P(\boldsymbol{C})$ を文字 n-gram 確率の積で近似したものである。すなわち

$$P(\boldsymbol{C}) \simeq \prod_{i=1}^{k} P(C_i|C_{i-(n-1)}\cdots C_{i-2}C_{i-1}) \quad (1)$$

2.2 単語 n-gram モデル

ここでいう単語とは、ある長さの文字の列である。単語集合は学習時に獲得され、学習時に出現した単語のみが存在するものとする。ただし、長さ 1 の単語は常に存在を仮定する。与えられた文字列$\boldsymbol{C}$は、単語の列 $w_i\cdots w_{i+j}$ に分割できる。この分割は一意ではない。ひとつの分割結果 $w_i\cdots w_{i+j}$ を$\boldsymbol{C}$のひとつの解析結果と呼び、$w_i\cdots w_{i+j} \in \boldsymbol{C}$ と書く。分割が一意でないことは、単語境界が曖昧であることを意味している。

単語 n-gram モデルは、$\boldsymbol{C}$の全ての解析結果に対する単語生起確率の積の総和であり、次式で与えられる。

$$P(\boldsymbol{C}) \simeq \sum_{w_1\cdots w_l \in \boldsymbol{C}} P(w_1\cdots w_l) \quad (2)$$

$$\simeq \sum_{w_1\cdots w_l \in \boldsymbol{C}} \prod_{i=1}^{l} P(w_i|w_{i-(n-1)}\cdots w_{i-1}) \quad (3)$$

さらに、簡単のためこの確率を最適な解析結果に対する単語生起確率の積で近似する。

$$P(\boldsymbol{C}) \simeq \max_{w_1\cdots w_l \in \boldsymbol{C}} \prod_{i=1}^{l} P(w_i|w_{i-(n-1)}\cdots w_{i-1}) \quad (4)$$

2.3 削除補間法

n-gram 確率を求めるためにはサンプル中に出現する n-gram の数をカウントすれば良いが、$n = 3$ と比較的小さな場合においてもその推定すべきパラメータ数はサンプルの量に比べて大き過ぎ、サンプルに対する過学習が起こってしまう。そこで、通常はこの影響を低減するために n-gram 確率にスムージングを導入する。

スムージングの方法の一つとして、より低次の n-gram 確率を併用し、全体の確率を複数のコンテキスト下での確率の重み付き線形和で表す方法がある。例えば、文字の trigram 確率は次のように bigram, unigram といった低次の確率と、zerogram 確率と呼ばれる定数とを併用して求める。(なお、右辺の確率は真の n-gram 確率ではなく、有限のサンプルから推定した値である。)

$$\hat{P}(C_i|C_{i-2}C_{i-1}) = \lambda_{c3}P(C_i|C_{i-2}C_{i-1}) + \lambda_{c2}P(C_i|C_{i-1}) + \lambda_{c1}P(C_i) + \lambda_{c0}P_{c0} \quad (5)$$

ただし $\lambda_{c3} + \lambda_{c2} + \lambda_{c1} + \lambda_{c0} = 1$.

削除補間法[2]は、このような複数の n-gram モデルを併用する場合の重み係数を、未知テキストの生成確率が近似的に最大となるように決定する手法である。削除補間法を利用してスムージングを行った日本語の文字 n-gram モデルは、サンプルが少数であったり、認識タスクが訓練サンプルと性質を異にする場合においても頑健である[1]。

2.4 文字・単語の複合 n-gram モデル

単語 n-gram モデルについても、文字の場合と同様に低次の単語 n-gram モデルを併用し、削除補間法を用いることによって少数サンプルの影響を低減することはできる。しかし、単語の場合にはこれでも不十分であると考えられる。その理由は次の通りである。

- 単語の種類は文字に比べてはるかに多く、少数サンプルの影響が深刻であるため
- 未知文章中に未知語 (モデルで想定していない単語) が存在することを避け得ないため

単語 n-gram モデルを少数サンプル下でも頑健に改善するために、これらの問題点がないモデルを併用することが考えられる。文字 n-gram モデルはまさにその要求を満たしており、これらのモデルを融合させることによって、より良いモデルを得ることができると期待される。

削除補間法は、一般に複数の異なる言語モデルの併用のために適用可能であると主張されている[2]。本報告で提案する複合 n-gram モデルもこの考えに基づいている。複合 n-gram 確率は、次式で定義される。

$$\hat{P}(w_i|w_{i-(n-1)}\cdots w_{i-1}) \equiv \lambda_w\hat{P}_w + \lambda_c\hat{P}_c \quad (6)$$

ただし $\lambda_w + \lambda_c = 1$. また、$\hat{P}_w$は低次の n-gram によって補間された単語 n-gram 確率であり、$\hat{P}_c$はコンテキスト $w_{i-(n-1)}\cdots w_{i-1}$ の下で単語 w_iが生起する確率を、低次の m-gram によって補間された文字 m-gram 確率の積によって求めたものである。すなわち、

$$\hat{P}_w \equiv \lambda_{w\,n}P_{w\,n} + \lambda_{w\,n-1}P_{w\,n-1} + \cdots + \lambda_{w0}P_{w0} \quad (7)$$

$$\hat{P}_c \equiv \prod_{j=1}^{l}\sum_{k=0}^{m}\lambda_{c\,k}P(C_{i,j}|C_{i,j-(k-1)}\cdots C_{i,j-1}) \quad (8)$$

ただし $\sum_k \lambda_{w\,k} = \sum_k \lambda_{c\,k} = 1$, $P_{w,k}$は単語 k-gram 確率、また $C_{i,j}$は w_iの j番目の文字である。このようにして定義した複合 n-gram 確率で式 (3) 中の単語 n-gram 確率を置き換えたものが、複合 n-gram モデルである。

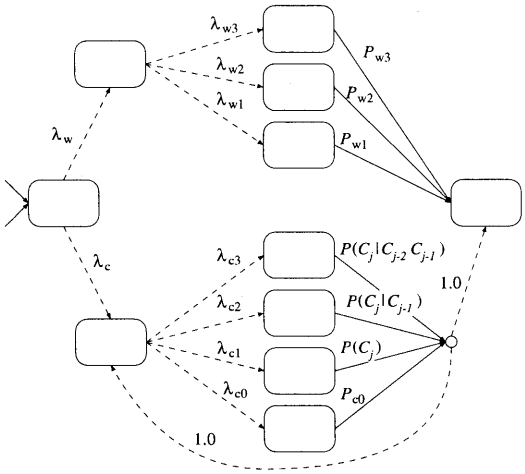

図 1. 複合モデルの状態遷移図。ここでは単語 trigram, 文字 trigram の例

複合 n-gram モデルの状態遷移図の例を図 1に示す。図中、点線の矢印はヌル遷移を示し、状態遷移系列が出力系列によって一意に決まらないという点で隠れマルコフモデルとみなすことができる。単語境界の曖昧性は、式 (4) を採用することによって無視することができるので、$P_{w,k}$は既知パラメータと考えて良い。したがって、再推定すべき未知パラメータはヌル遷移の確率、すなわち各モデルへの重みだけとなる。

3. モデルの学習法

3.1 単語 n-gram モデルの学習

本研究では、べた書きテキストから学習可能なモデルの構築を目的としているので、単語 n-gram の学習においては何らかの基準で訓練テキストを分割

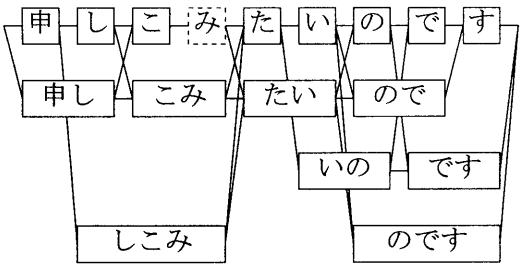

図 2. Viterbi アルゴリズムによる文の分割

し、単語列とみなす必要がある。

分割には、形態素解析による方法と、形態素解析によらずにテキストから何らかの基準で単語を抽出する方法が考えられる。後者については近年盛んに研究されている[3, 4]が、ここでは汎用形態素解析ツール JUMAN[5]を用いることとする。すなわち、訓練テキストを JUMAN によって形態素列に最適に分割し、その字面のみに注目して単語 n-gram の出現頻度を調べる。

3.2 削除補間法による重みの学習

削除補間法による重みの学習は、まず訓練テキストをいくつかの部分に分割し、ついで各々の部分に対して訓練テキストからの削除と各モデルの寄与度を求めるという手順で行われる。削除された部分 (今回は句点「。」で終わる列を文として文ごとに削除した) が仮想的に未知テキストの役割を果たす。このため、この部分のテキストについては分かち書きは行われていないものと仮定せねばならず、寄与度を求めるプロセスでは可能な単語列に対して Forward-Backward アルゴリズムを実行する必要がある。

今回は簡単のため、可能な単語列全てを考慮する代わりに、Viterbi アルゴリズムを適用して削除部分を単語列に分割し、その上で各モデルに対する寄与度を求め重みを修正するという方法を取った。Viterbi アルゴリズムによる文の分割の様子を図 2に示す。この図では、「申しこみたいのです」という文の解析を示している。矩形は単語 n-gram 確率テーブルに存在する単語を示す。点線の矩形は実際にはテーブルに存在しなかった単語のセルであるが、長さ 1 の単語についてはテーブルに存在するか否かを問わずセルを設けている。これは、どのような文に対しても正の確率を持つパスが存在することを保証したいためである。解析は、単語 unigram 確率テーブルをハッシュにより実現した擬似トライ構造[6]とすることで、効率良く行える。

4. 評価実験

提案した文字と単語の複合 n-gram モデルを評価するため、未知テキストに対するパープレキシティを求めた。モデルの訓練テキストには、新聞社説記事 (97–580 万字) を用いた。評価テキストとしては、訓練テキストと重複しないような新聞社説記事タスク (80098 字), および新聞の連載コラム記事タスク (23315 字) を用いた。

評価の基準に用いたのは、次式で推定される (テストセット) パープレキシティである。

$$PP \simeq \hat{P}(C_1C_2\cdots C_k)^{-1/k} \tag{9}$$

ただし、$\hat{P}$は評価する言語モデル、$C_1C_2\cdots C_k$は評価テキストである。パープレキシティは認識の困難さを示す値であり、同一タスクにおいては小さいほど優れた言語モデルと言える。式 (9) の右辺の確率を求めるためには、評価テキスト全文に対して図 2 のような文の分割が必要である。

評価した言語モデルは次の通りである。

文字モデル 式 (6) において$\lambda_c = 1$ とした場合に相当する。

単語モデル 式 (6) において$\lambda_w = 1$ とした場合に相当する。なお、式 (7) 中の単語 zerogram 確率 P_{w0} は JUMAN の語彙数の逆数により与えた。

(線形補間型) 複合モデル 式 (6) 中の重み係数をも削除補間法によって求めたものである。式 (7) 中の P_{w0} は、文字 n-gram モデルと組み合わせたことによって不要になるため 0 とした。

back-off 型複合モデル 提案した削除補間法に基づく複合モデルと比較するためのものである。back-off スムージング法[7]とは、信頼性の低い n-gram 確率を、より低次の n-gram 確率で置き換えることによって平滑化する手法である。通常は unigram が信頼できない場合には zerogram に back-off するが、今回は単語 zerogram 確率の代わりに式 (8) を使って文字 n-gram に back-off することを試みた。これによって、削除補間法に基づくモデルと同様に、少数サンプルや未知語の影響が軽減されると考えられる。

back-off の方法としては、以下を用いた[8]。単語を w, 単語の出現したコンテキストを x とし、コンテキスト x で w が出現する回数を $c_{w|x}$, コンテキスト x で出現する単語の数を n_x, コンテキスト x で出現する単語の種類を r_x としたと

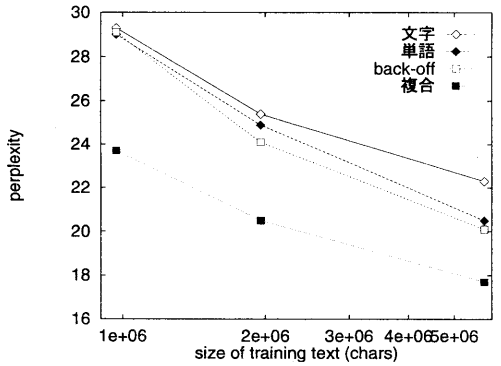

図 3. 新聞社説タスクに対するパープレキシティ

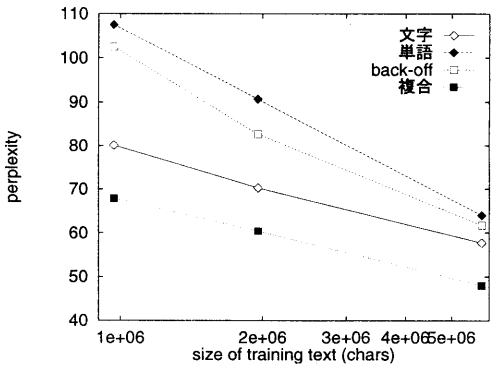

図 4. コラム記事タスクに対するパープレキシティ

き、単語 n-gram 確率を次式で求める。

$$\hat{P}(w|x) = \begin{cases} \dfrac{c_{w|x}}{n_x + r_x} & \text{if } c_{w|x} > 0 \\ \dfrac{r_x}{n_x + r_x}\hat{P}(w|x') & \text{if } c_{w|x} = 0 \text{ and } n_x > 0 \\ \hat{P}(w|x') & \text{if } n_x = 0 \end{cases} \quad (10)$$

ここで x' は x より 1 つ低次のコンテキストである。また unigram の場合には特別に、

$$\hat{P}(w) = \begin{cases} \dfrac{c_w}{n_x + r_x} & \text{if } c_w > 0 \\ \dfrac{r}{n + r}\hat{P}_c & \text{otherwise} \end{cases} \quad (11)$$

とする。

これら全てについて、文字・単語の両方とも trigram モデルを用いた。

図 3に、新聞社説タスクに対するパープレキシティを示す。横軸は訓練テキストの量である。図中、単に「複合」とあるのは削除補間法に基づく線形補間型複合モデルである。複合モデルの性能が他のものに比べ際だっていることがわかる。文字モデルは訓練テキストを増やした時の性能の上昇が緩やかである。back-off 型モデルは、単語モデルの場合とほとんど変わらない。

図 4に、コラム記事タスクに対するパープレキシティを示す。このタスクは訓練テキストとあまり性質が似ていない。社説タスクの場合に比べ、単語モデルと back-off 型モデルの性能の低下が著しい。これは、共通する語彙が少なくなったためと考えられる。この場合においても、複合モデルは優れている。これは、頻出単語については単語モデルを、未知語については文字モデルをといったように互いの良い部分をうまく引き出せた結果と考えられる。また back-off 型モデルの場合に、unigram 確率が単語モデルに back-off される機会はあまりないことが読み取れる。

5. あとがき

文字 n-gram と単語 n-gram を融合させた、べた書きテキストから学習可能な言語モデルの構成法を示し、提案したモデルが他のものに比べ優れていることを実験により示した。

今回は単語知識の獲得に形態素解析システムを用いたため、真に訓練テキストのみからモデルを形成したものとは言えない。また、この枠組では形態素解析システムがあらかじめ持っている形態素辞書にない単語を学習によって獲得することができない。今後は、形態素辞書や文法などの知識への依存を減らし、字面処理を主とした言語モデル獲得の方法について検討していきたい。

謝辞

JUMAN を開発され、またフリーソフトとしてご提供くださっている京都大学工学部長尾研究室および奈良先端科学技術大学院大学松本研究室の皆様に感謝いたします。

参考文献

[1] H. Mori, H. Aso, and S. Makino, "Robust n-Gram Model of Japanese Character and Its Application to Document Recognition," IEICE Trans. Inf. & Syst., vol.E79-D, no.5, pp.471–476, 1996.

[2] 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.

[3] 伊藤, 好田, "文字列パターンの N-gram による文節モデルの検討", 信学技報, NLC95-61, 1995.

[4] 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.

[5] 松本, 黒橋, 宇津呂, 妙木, 長尾, "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.

[6] 中村, 黒橋, 長尾, "部分文字列情報の利用による日本語単語の高速検索", 情報処理学会研究報告, NL-101-12, 1994.

[7] S.M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," IEEE Trans., vol.ASSP-35, no.3, pp.400–401, 1987.

[8] P. Placeway, R. Schwartz, P. Fung, and L. Nguyen, "The Estimation of Powerful Language Models from Small and Large Corpora," Proc. ICASSP-93, vol.II, pp.153–176, 1994.