# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| 論題(和文) | N-gramカウントを用いた言語モデルの効率的な選択学習 |
| Title(English) | |
| 著者(和文) | 久保田 雄, 篠崎 隆宏, 古井 貞熙, 宇都宮 栄二, 新堂 安孝 |
| Authors(English) | Yu Kubota, Takahiro Shinozaki, SADAOKI FURUI, Eiji Utsunomiya, Yasutaka Shindo |
| 出典(和文) | 日本音響学会2011年春季講演論文集, , No. 3-5-2, pp. 73-74 |
| Citation(English) | , , No. 3-5-2, pp. 73-74 |
| 発行日 / Pub. date | 2011, 3 |

# N-gram カウントを用いた言語モデルの効率的な選択学習*

☆久保田雄, 篠崎隆宏, 古井貞熙 (東工大), 宇都宮栄二, 新堂安孝 (KDDI 研)

## 1 はじめに

連続音声認識において高い認識率を得るためには，認識対象タスクに適合した大量のデータを用いて言語モデルを学習する必要がある．しかし，タスク毎に大量の学習データを作成することは非常にコストがかかり非現実的である．そのため実際には認識タスクとは異なる既存のコーパスを用いてモデル学習を行う．その際目的タスクの認識に適した言語モデルを得るには，目的タスクに近い適切なサブセットを選択し学習に用いることが重要であると考えられる．

既存の手法として，少量の目的タスクテキスト (開発データ) からモデルを構築し学習データの各文に対するパープレキシティを求め，閾値より値の低い文を選択する手法が提案されている [1] [2]．これにより目的タスクに登場しやすい表現を多く含んだ文を優先的に選ぶことができる．しかしこの手法は開発データからモデルを学習して学習データに対する言語尤度最大化を行っており，直接的に開発データに対する言語尤度を最大化していない点が問題と考えられる．そこで逆に学習データから学習した言語モデルの開発データに対する言語尤度を最大化するような学習データ選択のアプローチが考えられる．しかしこの方法は計算コストの多さが問題となり，これまで研究が行われてこなかった．本研究では N-gram カウントを用いることでその計算量を削減した，順向き最尤基準選択法を提案する。

## 2 順向き最尤基準選択法

提案する順向き最尤基準選択法 (Direct-Likelifood-Maximization Selection) のプロセスを図 1 に示す．まず学習データ全体から言語モデル $M(0)$ を学習する．そしてそのモデルの開発データに対するパープレキシティを $PP(0)$ とする．次に学習データ全体を $L$ 文ずつのブロック $T(1), T(2), \ldots, T(K)$ に分割する．今 $T(1)$ のみを除いた残りの学習データから $M(\bar{1})$ を学習し、$PP(\bar{1})$ を求める。もし $PP(\bar{1}) - PP(0) < 0$ ならば $T(1)$ を取り除いた方がパープレキシティが小さくなるということなので，$T(1)$ は取り除いた方が良い．逆に $PP(\bar{1}) - PP(0) > 0$ ならば $T(1)$ は学習データとして用いた方がよいと判断する。そして $T(1)$ を再び学習データに戻した上で $T(2), \ldots, T(K)$ について

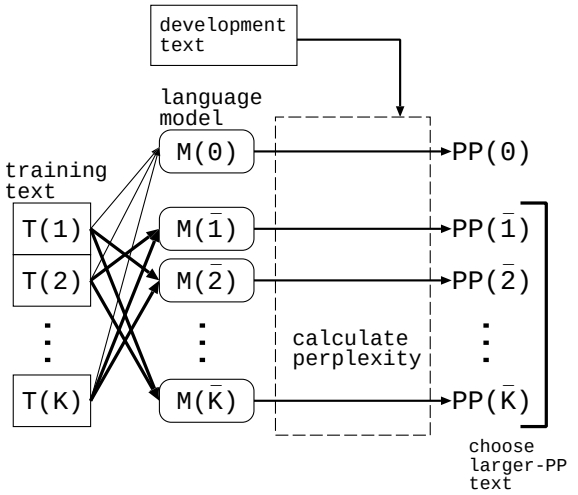


Fig. 1 提案法の文選択アルゴリズム

も同様の作業を行い、用いた方が良いと判断されたブロックをまとめることで最終的な選択データとする。なおここで適当な閾値 $\alpha$ を用いて $PP(\bar{i}) - PP(0) > \alpha$ となるテキストを選択することも可能である．

このアプローチにおいて毎回テキストから言語モデルを学習することは非常に計算コストがかかる。そこで学習データの N-gram カウントを用いることで，計算量の削減を図る．開発データ ( $w_1^n$ ) の出現確率は学習データ中の N-gram カウント $c(w)$ を用いて

$$P(w_1^n) = \prod_{j=1}^{n} P(w_j | w_{j-N+1}^{j-1}) = \prod_{j=1}^{n} \frac{c(w_{j-N+1}^{j})}{c(w_{j-N+1}^{j-1})} \quad (1)$$

となる．ここであるブロック T(i) に含まれている N-gram の出現回数を一般に $subc_{(i)}(w)$ で表す．するとそのブロックを取り除いた場合の開発データの出現確率は

$$P_{(i)}(w_1^n) = \prod_{j=1}^{n} \frac{c(w_{j-N+1}^{j}) - subc_{(i)}(w_{j-N+1}^{j})}{c(w_{j-N+1}^{j-1}) - subc_{(i)}(w_{j-N+1}^{j-1})} \quad (2)$$

と再計算できる．このように計算すればサブセットを取り除く度に最初から言語モデル全体を学習し直す必要がなく，計算量を大幅に削減することができる．なお，バックオフを考慮すると計算量が増えるため、学習データ中に存在しない N-gram の出現確率は (N-1)-gram の出現確率で近似する．また計算量はテキストの分割数 $K$ に比例して増えるが，$K$ が大きいほど より細かい選択が行えるというトレード オフが存在する．

*Efficient selective language model training using N-gram count. by Yu Kubota, Takahiro Shinozaki, Sadaoki Furui (Tokyo Institute of Technology), Eiji Utsunomiya, and Yasutaka Shindoh (KDDI R&D Labs)

実際には開発セットに対する尤度の最大化のみを考えることで不利益が発生する場合がある．開発セットに対する尤度を目的関数とする場合，開発セットに含まれない N-gram は全て捨てた方が評価値が向上する．しかし，開発セットの大きさは有限であることから，このままでは開発セットにたまたま出現しなかった認識に有用な N-gram を捨てすぎてしまうことが考えられる．そこで，開発セットと関連の高い N-gram を含むブロックが優先的に保存されるよう，式 3 に示すコンテキスト局在度重み (Context Locality Weight:CLW) を提案し，目的関数への導入を試みた．

$$\hat{P}_{(i)}(w_j|w_{j-N+1}^{j-1}) = P_{(i)}(w_j|w_{j-N+1}^{j-1})\left(1-\frac{subc(w_{j-N+1}^{j-1})}{c(w_{j-N+1}^{j-1})}\right) \quad (3)$$

ここで $subc()$ は取り除いたブロック内に存在する N-gram の出現回数である．CLW はあるブロック内に特定の (N-1)-gram が特異的に存在するとき，その (N-1)-gram をコンテキストとして持つ全ての N-gram 確率を割り引くように働く．その結果，開発セットと関連が高く貴重なブロックを保存しやすくなる．

## 3 実験条件

認識には東京工業大学で開発を行っている WFST 音声認識デコーダー (T$^3$decoder)[3] を用いた．言語モデルの学習データは Yahoo!ブログ [4] の投稿記事を形態素解析したテキストデータ (形態素数:555M 語，25M 文，語彙数:972K 語) である．使用した形態素解析器は MeCab であり，先行研究 [5] を参考に開発データを用いたフィラーの挿入を行った．言語モデル生成時には出現回数が 3 回に満たない 2-gram，3-gram をカットオフし，出現頻度上位 30K 語についてモデル化した．開発データには CSJ の模擬講演書き起こしデータ 100 講演分 (形態素数:235K 語，14K 文，語彙数:12K 語) を用いた．音響モデルは日本語話し言葉コーパス (CSJ) の学会講演音声 254 時間より EM 学習した 3000 状態 32 混合の状態共有トライフォンモデルである．特徴量は MFCC12 次元と対数エネルギー，及びそれらのデルタ項，デルタデルタ項の計 39 次元である．評価セットは異なる話者男女 10 名による模擬講演 10 講演からなる CSJ 評価セットである．

## 4 実験結果

図 2 は選択した学習データで言語モデルを構築した場合の音声認識の結果であり，横軸は閾値 $\alpha$ を変えることで選択した学習データの量を，縦軸は単語誤り率を示す．グラフ中で rand はランダム選択である．また baseline は開発データから言語モデルを学習し，

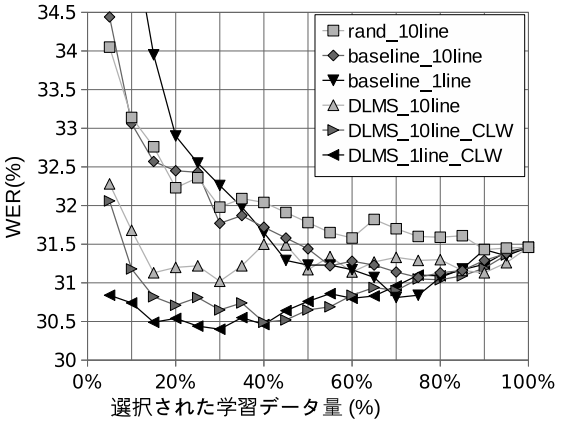

Fig. 2　選択したデータ量と認識性能

学習データの各ブロックに対して求めたパープレキシティが低いブロックを優先的に選ぶ従来法である．そして DLMS が提案法である．$X$line は学習データ分割時の 1 ブロックあたりの文数を表し，CLW は提案手法で CLW の補正項を加えた場合を示す．

10 文単位で文の選択を行った場合の最低単語誤り率を比較すると，従来法ではランダムの 31.4%よりも低い 31.1%となった．CLW を加えない提案法は 31.0%であり，従来法よりも小さい言語モデルで同程度以下の単語誤りを得られた．そして CLW を加えた提案法ではさらに低い 30.5%となった．

1 文単位で選択する場合，単語誤り率の最小値は従来法が 30.8%だったのに対し，CLW を加えた提案法では 30.4%と従来法よりも低かった．また従来法では単語誤り率最小時 (選択データ量:70%) に言語モデル内の 3-gram が 11.7M 種類であったのに対し，提案法では同程度の単語誤り率時 (選択データ量:5%) に 2.8M 種類と従来法の 1/4 程度となった．

## 5 まとめ

開発データに対する言語尤度が最大となるような言語モデルを構築するための学習データの選択学習手法を提案した．実験の結果，提案法は従来法よりも高い認識性能を得ることができ，かつモデルサイズの縮小化にも有効であると示した．

## 参考文献

[1] Ryuichi Nishimura 他，Eurospeech，pp.2127-2130， 2001.
[2] 翠輝久 他，電子情報通信学会技術研究報告 106(442)，pp.67-72，2006-12-15.
[3] P.R.Dixon 他，IEEE ASRU，443-448，2007.
[4] http://blogs.yahoo.co.jp/
[5] 古井貞熙 他，ICSLP2000 Beijing，pp.518-521，2000.