2D-7

# 外部知識としてウェブを用いた3-gram言語モデル拡張手法の検討

西村 竜一　島田 敏明　田中 雅康　河原 英紀　入野 俊夫
和歌山大学 システム工学部

## 1 はじめに

我々は、大語彙連続音声認識システムの初期開発コストを抑えるために、少量なテキストから学習した言語モデルの活用に着目している。一般に、言語モデルの作成には、新聞記事10年分以上の量を持つテキストが必要であり、その準備が音声認識システムの開発コストを上げる要因であった。本研究の目的は、比較的小規模な言語モデルを用いて、従来と同程度の音声認識精度を得ることである。本稿では、その実現を目指し、ウェブから取得した外部知識を用いた単語3-gram言語モデルの拡張技術について検討したので報告する。

## 2 単語3-gramモデルにおける未観測問題

音声認識は、入力信号系列 $X$、出力単語列 $W$ ($w_1^k = w_1 w_2 \cdots w_k$) としたとき、

$$W = \arg\max_W P(X|W)P(W) \tag{1}$$

によって、最尤解 $W$ を求める問題である。ここで、$P(X|W)$、$P(W)$ を与えるのが音響モデル、言語モデルであり、本研究では言語モデルのみを取り扱う。

言語モデルには、一般に単語3-gramモデルを用いる。このとき、下式で単語列の出現確率 $P(W)$ を近似する。

$$P(W) = \Pi_{i=1}^k P(w_i|w_{i-1}w_{i-2}) \tag{2}$$

このとき、条件付き確率 $P(w_i|w_{i-1}w_{i-2})$ は、学習元テキスト中の単語列 $w_{i-2}^i$ の出現頻度 $C_{i-2}^i$ を用いて、下記の最尤推定の式により算出できる。

$$P(w_i|w_{i-1}w_{i-2}) = \frac{C_{i-2}^i}{C_{i-2}^{i-1}} \tag{3}$$

式(3)において、システム語彙辞書の単語の組み合わせである3-gram $w_{i-2}^i$ が学習元に出現しない場合、未観測3-gramとなる。3-gramの未知状態を防ぐため、単語3-gramモデルの作成には大量のテキストを要する。ただし、大語彙システムでは、テキストを増やしても未観測3-gramは必然的に発生する。そのため、未観測となった3-gramの確率を、既知の2-gramの確率から推定するバックオフ平滑化法[1, 2]が広く用いられる。

## 3 提案手法

提案手法は、バックオフ平滑化手法と併用することで、未観測3-gramの確率値を推定し、音声認識精度の向上を得るものである。バックオフ平滑化は、内包的な手法であるのに対し、本手法は、ウェブリソースから抽出した外部情報を吸収し、単語3-gramモデルを拡張する。拡張された3-gramモデルでは、保持する3-gramが追加されており、バックオフの発生が抑えられる。

今回、ウェブリソースには、Google社の「Web日本語Nグラム第1版（Google N-gram）」[3]を用いた。式(3)において、$C(w_{i-2}^i) = 0$ となった未観測3-gram $w_{i-2}^i$ に対し、その出現頻度の相当数 $\widehat{C}(w_{i-2}^i)$ をGoogle N-gramの登録情報に基づき算出する。

1. 学習元テキストから抽出した2-gram $w_{i-2}^{i-1}$ を既知とし、単語を1つ接続した3-gram $w_{i-2}^i$ を構成する。
2. $w_{i-2}^i$ のうち、$C(w_{i-2}^i) = 0$ のものを未観測3-gram $w_{i-2}^i$ とする。
3. すべての未観測3-gram $w_{i-2}^i$ に対して、Google N-gram内に登録されているその出現頻度を $C_{google}(w_{i-2}^i)$ とする。
4. 下式により学習元テキストとGoogle N-gramのスケールの違いを調整した $\widehat{C}(w_{i-2}^i)$ を求める。

$$\widehat{C}(w_{i-2}^i) = C_{google}(w_{i-2}^i) \times \frac{N_{orig}}{N_{google}} \times \alpha \tag{4}$$

$N_{orig}$ は、学習元テキストの3-gram頻度の総和、$N_{google}$ は、Google N-gramの3-gram頻度総和である。$\alpha$ ($\alpha < 1$) は、スケール調整係数である。

5. $\widehat{C}(w_{i-2}^i) \geq 1$ のとき、既知の $C(w_{i-2}^{i-1})$ 及び $\widehat{C}(w_{i-2}^i)$ を式(3)に適用し、$P(w_i|w_{i-1}w_{i-2})$ を算出する。この結果、未観測であった3-gram $w_{i-2}^i$ はモデルに追加される。一方、$\widehat{C}(w_{i-2}^i)$ が1未満のときは、追加から除外される。

Google N-gramに登録済みの3-gramは、学習元テキストから抽出できる3-gramよりも圧倒的に多い。このため、Google N-gramから獲得できた3-gramをすべて用いて拡張すると、拡張後のモデルから学習元テキストの特徴が失われる可能性が高い。提案手法では、$\alpha < 1$ を用いて調整することで、それを制限することにした。

また、$w_{i-2}^i$ の中に重要単語が1つも含まれていない場合は、$\alpha = 0$ とし、追加から除外した。これによって、認識対象に対して関連の高いと想定される3-gramのみを選別し、追加することが可能になった。このとき、重要単語は、TF-IDF及びYahoo!関連検索ワードWeb API[4]をもとに名詞のみを抽出して定義した[5, 6]。

## 4 評価実験

実験により、提案手法で拡張した単語3-gramモデルの音声認識性能をJulius 4.1.5[7]を用いて評価した。本実験では、学習元テキスト及び評価用テストセットに「日本語話し言葉コーパス（CSJ）」[8]を用いた。「政治」を認識対象のトピックとし、CSJの中から「政治」という単語が一つでも含まれる講演テキストを対象にした。

**A Proposal of Expanding 3-gram Language Model Based on External Web Knowledge**, Ryuichi NISIMUR, Toshiaki SHIMADA, Masayasu TANAKA, Hideki KAWAHARA, Toshio IRINO (Faculty of Systems Engineering, Wakayama University)

表 1: 拡張前単語 3-gram モデルの仕様

| | |
|---|---|
| 認識トピック | 政治 |
| 学習元テキスト | CSJ[8] 140 講演 |
| 総文書数 | 9,331 |
| 総単語数 | 414,960 |
| 3-gram エントリ数 | 250,702 |
| Back-off 手法 | Witten-Bell[2] |

表 2: 比較実験結果（単語誤り率 [%]）

| | バックオフ | 拡張 | 単語誤り率 |
|---|---|---|---|
| (1) | × | × | 60.79 |
| (2) | × | ○ | 60.67 |
| (3) 従来モデル | ○ | × | 27.73 |
| (4) 併用モデル | ○ | ○ | **26.09** |
| (5) Google 全て | ○ | - | 44.83 |

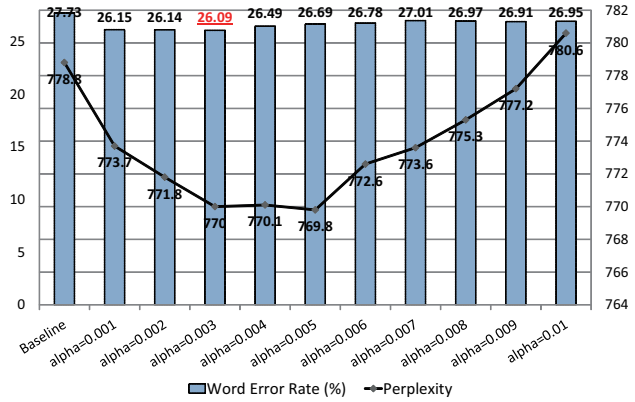


図 1: 実験結果（単語誤り率及びテストセットパープレキシティ）

表 1 に、提案法による拡張前の単語 3-gram モデルの仕様を示す。ここで、学習元テキストの量は、通常の音声認識用言語モデルの作成時より少なく調整した。テストセットは、男性 6 名、女性 2 名による「政治」に関する 8 講演音声（学会講演 2, 模擬講演 6, 総文書数 636, 総単語数 16,072）である。

実験結果として、単語誤り率（Word Error Rate）及び 3-gram テストセットパープレキシティを図 1 に示す。図中の **Baseline** は、拡張前モデルの性能を示す。式 (4) の調整係数 $\alpha$ には、$0.001 \sim 0.01$ の値を用いた。

実験結果より、単語誤り率 26.09% ($\alpha = 0.003$) を得ることができた。これは、**Baseline** と比較して、1.64 ポイントの精度改善であり、有意差を確認できた。パープレキシティに関しても、拡張を適用することにより値の低下、つまり、性能改善を確認することができる。以上の結果より、提案手法の有効性を確認することができた。ただし、$\alpha$ の値を大きくする（$\alpha = 0.005$ 前後）と、性能劣化に転じる。これは、認識対象に関係の小さい 3-gram が無用に過追加されたことにより、言語モデルが音声認識の最尤解探索に与える情報量が減少したことが原因と考えられる。この結果は、$\alpha$ の決定方法が重要であることを意味している。

加えて、参考のために下記 5 条件の比較を行った。

(1) バックオフ無し、拡張無し
(2) バックオフ無し、拡張適用 $\alpha = 0.003$
(3) バックオフ有り、拡張無し（従来モデル）
(4) バックオフ有り、拡張適用 $\alpha = 0.003$（併用モデル）
(5) Google N-gram すべてを用いたモデル、バックオフ平滑化有り

このうち、(3) が一般的な従来モデル、(4) が本研究の併用モデルとなる。また、(5) は、学習元テキストを使わずに、Google N-gram に登録済みの 3-gram のみで作成したモデルである。Google N-gram には、3.93 億個の 3-gram エントリがあり、大規模なモデルである。

表 2 に単語誤り率を示す。バックオフ平滑化が有効であることを改めて確認する結果となった。この結果からバックオフ平滑化と提案手法の併用が有用であると考えることができる。

## 5 まとめ

本稿では、Google N-gram から取得した頻度情報に基づく大語彙連続音声認識用単語 3-gram モデルの拡張手法を提案、実験で評価した。本手法は、追加のテキストの収集無しに、比較的簡単な処理で音声認識の性能を向上できる有効な手法である。また、従来からのバックオフ平滑化との併用が有用であることを確認した。

今後の課題としては、$\alpha$ の値の調整方法や、本稿では詳細を省いた、認識対象と関連性の高い重要単語の決定手法の検討 [5, 6] が挙げられる。

**謝辞** 本研究は、(独) 科学技術振興機構 研究成果最適展開支援事業 A-STEP フィージビリティスタディ (FS) 探索タイプ及び和歌山大学 H22 年度学長裁量経費を受けて実施した。

### 参考文献

[1] Katz, S., Estimation of probabilities from sparse data for the language model component of a speech recogniser, *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol.35, no.3, 400–401, 1987.
[2] Witten, I.H., The Zero-Frequency Problem: Estimating The Probabilities of Novel Events in Adaptive Text Compression, *IEEE Trans. Information Theory*, vol.37, no.4, pp.1086–1094, 1991.
[3] 工藤 拓, 賀沢 秀人, Web 日本語 N グラム第 1 版, 言語資源協会, 2007.
[4] http://help.yahoo.co.jp/help/jp/search/web/web-17.html
[5] 島田 敏明 他, 単語重要度を用いた N-gram 補完手法が与える音声認識性能の調査, 情報処理学会研究報告, 2010-SLP-82-19, 2010.
[6] 島田 敏明 他, 講演発話を用いた N-gram 補完手法の音声認識性能評価, 日本音響学会 2010 年秋季研究発表会講演論文集, pp.147–148, 2010.
[7] Lee, A., et al., Julius – An Open Source Real-Time Large Vocabulary Recognition Engine, Proc. *Eurospeech2001*, pp.1691–1694, 2001.
[8] Maekawa, K., Corpus of Spontaneous Japanese: Its design and evaluation, Proc. *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 7–12, 2003.