3C-3

ウェブデータベースを用いた音声認識用言語モデルの簡易適応

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

前回報告 [1] したように、我々は、Google N-gram を用いた音声認識用 3-gram モデルの拡張手法を検討している。同手法より、限られた量のテキストから学習した言語モデルを用いて、従来の一般的な手法と同程度の音声認識精度を得ることに成功した。言語モデル学習用のテキストは、大量に必要であり、事前準備に手間を要するため、音声認識システム全体の開発コストを引き上げる要因であった。本研究によって、その量を削減できるようになった。本稿では、提案法を用いて、学習元となるテキストとは異なるタスクの言語モデルを生成するトピック簡易適応について検討した。

2 3-gram モデルと提案法

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に着目する。

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

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

右辺の条件付き確率 $P(w_i|w_{i-1}w_{i-2})$ は、単語列（3-gram）w_{i-2}^i の出現頻度 $C(w_{i-2}^i)$ から算出する。

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

通常の手順では、形態素解析を適用し、単語分割された学習元のテキスト T の 3-gram w_{i-2}^i を数え上げて、頻度 $C(w_{i-2}^i)$ を得ることができる。学習元テキスト T には、通常、新聞記事数年分に相当する量が要求される。

ここで、十分な量のテキスト T を事前に用意することは困難な場合を想定する。通常手順で学習した 3-gram モデルの言語的な特徴（受理可能な語彙や言い回し等）は、元のテキストを引き継ぐ。このため、ある特定のトピックを受理可能なモデルを構築するには、そのトピックに関連する内容が記述されたテキストを収集し、整備する必要があった。しかし、その整備には労力を要するため、システム開発の現場では、テキスト無し、もしくは、少量で、任意のトピックに対応した 3-gram モデルを生成する手法が求められている。

Building Method for Topic-Adapted Language Model Using Web Database for Automatic Speech Recognition Systems, Ryuichi NISIMURA, Toshiaki SHIMADA, Masayasu TANAKA, Hideki KAWAHARA, Toshio IRINO (Faculty of Systems Engineering, Wakayama University)

本研究では、通常手順で得た頻度 $C(w_{i-2}^i)$ に加え、ウェブの外部知識から抽出した頻度 $\widehat{C}(w_{i-2}^i)$ を用いて学習することで、望みのトピックに対応した 3-gram モデルの生成を実現する。外部知識としては、Google 社の「Web 日本語 N グラム第 1 版（Google N-gram）」[2] を用いた。手順を以下に述べる。

1. 提案法を適用する前の 3-gram モデルの学習元テキスト T から抽出できる単語のリストを L とし、3-gram w_{i-2}^i 及び頻度 $C(w_{i-2}^i)$ を準備する。
2. 続いて、開発者が望む適応先トピックに関連する単語を列挙したリストを $\widehat{L}$ として準備する。なお、今回の実験では、適応先トピックに関連する内容を含んだ、小規模なテキスト $\widehat{T}$ を別途用意し、それを構成する単語のリストを作成し、$\widehat{L}$ とした。
3. $\widehat{L}$ と L を合わせた語彙で構成する 3-gram w_{i-2}^i に対して、Google N-gram 内に登録されているその頻度を $C_{google}(w_{i-2}^i)$ とする。
4. 下式により、T と Google N-gram の総登録データ数のスケールの違いを調整した $\widehat{C}(w_{i-2}^i)$ を求める。

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

   N は、T の 3-gram 頻度の総和、N_{google} は、Google N-gram の 3-gram 頻度総和である。
5. $C(w_{i-2}^i)$ が存在せず、かつ、$\widehat{C}(w_{i-2}^i) \geq 1$ のとき、$\widehat{C}(w_{i-2}^i)$ を 3-gram モデルの学習に追加することとし、式 (3) より、$P(w_i|w_{i-1}w_{i-2})$ を算出する。$\widehat{C}(w_{i-2}^i)$ が 1 未満のときは、追加から除外される。
6. $\widehat{L}$ の単語と付随する 3-gram が取り込まれることによって、$\widehat{T}$ に関連する文章を受理できる 3-gram モデルが新たに生成される。

2.1 追加単語の選別

Google N-gram には、極めて多くの 3-gram の情報が含まれているため、$\widehat{L}$ の単語をすべて追加した 3-gram モデルは不要な情報を大量に含むことになる。これは認識精度劣化の原因となる [1]。今回は、追加する単語の選別し、追加される 3-gram を抑制することで、この問題を防ぐ。選別は 2 段階で処理した。

1: TF-IDF を用いた特徴語の抽出

TF-IDF[3] を用いて、$\widehat{T}$ から、特徴的な単語である 1,000 個（W_{tfidf}）を抽出した。

2: 3-gram 追加による情報量増減を基準とした選別

W_{tfidf} をさらに選別し、追加によって適応先のトピックに有用な情報を与える単語のリストを作成した。この際、単語の追加によって生じるパープレキシティの増減値を選別の基準とした。

表 1: テキスト諸元: 学習元 T, 適応先 $\widehat{T}$, 評価用 T_E

学習元 T	「政治」に関する 140 講演
総文数	9,331
総単語数	414,960
適応先 $\widehat{T}$	「野球」に関する 118 講演
総文数	6,082
総単語数	258,955
評価用 T_E	「野球」に関する 9 講演
総文数	516
総単語数	19,004

パープレキシティは、一般に、3-gram モデルの評価に用いられる指標である。単語数 n のテキスト T' に対する単語あたりのエントロピーは、

$$H(T') = -\frac{1}{n}\sum_{i=1}^{n} P(w_i|w_{i-1}w_{i-2}) \log P(w_i|w_{i-1}w_{i-2}) \quad (5)$$

であり、

$$PP = 2^{H(T')} \quad (6)$$

をパープレキシティと呼ぶ。 値が小さいとき、文章 T' が生成される確率が高く、その 3-gram モデルは、T' に対して高い性能を持つと言える。

本研究では、T' を適応先テキスト $\widehat{T}$ とした。W_{tfidf} のうち、ある一つの単語 w' を候補単語とする。w' を含む 3-gram の頻度を Google N-gram から抽出し、学習元の $C(w_{i-2}^{i})$ に追加し、w' 追加後の 3-gram モデルを学習する。w' の追加によって、$\widehat{T}$ に対するパープレキシティが改善（減少）した場合、w' を有用として選定する。これを W_{tfidf} のすべての単語に対して繰り返し、最終的に追加する有用単語のリストを決定した。

3 評価実験

提案法で生成した 3-gram モデルの音声認識性能を Julius 4.2[4] を用いて評価した。「日本語話し言葉コーパス（CSJ）」[5] の講演音声データを用いた。手法適用前の学習元テキスト T には、「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の書き起こし文章を用いた。適応先の $\widehat{T}$ は、「野球」という単語が含まれる講演文章を同じく CSJ から抽出した。評価では、男性 6 名、女性 3 名による「野球」に関する 9 講演（学会講演 4, 模擬講演 5, 総文数 516, 総単語数 19,004）の音声を認識し、単語誤り率を求めた。なお、評価で用いた前述 9 講演の書き起こし文章 T_E は、T 及び $\widehat{T}$ から除外している。表 1 に各テキストの諸元を示す。

3-gram モデルは、Palmkit[1] を用いて学習し、Witten-Bell 手法 [6] によるバックオフ平滑化を適用した。

図 1 は、前述した 2 段階の選別の結果、有用単語として選定した単語の一部抜粋である。今回、「野球」のトピックに対して、49 個の単語が有用と選別される結果となった。一方で、TF-IDF で抽出した W_{tfidf} の残りは、有用単語から除外することができており、おおよそ有効な選別を実現することができた。

巨人　長嶋　予選　野球　バッティング
虎　外野　球団　金メダル　バント　打者

図 1: 選別後の有用単語一部抜粋（野球）

表 2: 単語誤り率 [%]（「野球」9 講演を認識）

適用前: T から作成「政治」モデル	41.2
提案法: T を $\widehat{T}$ で適応	39.6
参考比較: $\widehat{T}$ から作成「野球」合致モデル	43.8

音声認識実験の結果を表 2 に示す。数値は単語誤り率であり、値が小さい方が高い精度を示す。提案法によって、学習元 T を $\widehat{T}$ で適応して学習した 3-gram モデルが 1.6%の精度向上を得ており、有意差を確認することができた。参考比較の $\widehat{T}$ を用いて作成した「野球」の合致モデルよりも高い精度となった。これは $\widehat{T}$ が T の 6 割程度の量であったため、3-gram モデルの素材のテキストとしては劣っていたことに原因がある。

4 まとめ

本研究では、Google N-gram を利用する 3-gram 拡張手法を応用し、既存の学習用テキスト「政治」とは異なるトピック「野球」に適応した 3-gram モデルの構築手法を検討した。また、「野球」に関連する比較的小規模の文章を用いて、追加する有用単語の選別を行い、膨大な無用 3-gram の追加による認識性能低下を防ぐことができた。実験により精度の向上を確認した。

今後は、さらなる精度向上を目指し、より詳細な検討が必要であると考える。また、プログラムを整理し、システム開発者が望むトピックに対応した 3-gram モデルを手軽に構築できるツールとして整備する予定である。

謝辞 本研究の一部は、和歌山大学平成 23 年度独創的研究支援プロジェクト及び科学研究費補助金の支援を受けた。

参考文献

[1] 西村 他, 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, 情報処理学会第 73 全国大会講演論文集, vol.2, pp.75–76, 2011.
[2] 工藤, 賀沢, Web 日本語 N グラム第 1 版, 言語資源協会, 2007.
[3] 天野 他, IT Text 自然言語処理, p.138, オーム社, 2007.
[4] Lee, A., et al., Julius – An Open Source Real-Time Large Vocabulary Recognition Engine, Proc. *Eurospeech2001*, pp.1691–1694, 2001.
[5] Maekawa, K., Corpus of Spontaneous Japanese: Its design and evaluation, Proc. *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 7–12, 2003.
[6] 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.

[1] http://palmkit.sourceforge.net/