言語モデルへのフィラー追加による講演音声の認識率向上

三本木　尚志　　浦谷　則好
東京工芸大学大学院　電子情報工学専攻

1. はじめに

音声認識では言語モデルと使用環境との適合度により認識率が大きく左右される．高い認識精度を得るには環境に適合する言語モデルが必要であるが，実際には言語モデルの作成に必要な書き起こし文のデータを獲得できることは少ない．また，書き起こし文の用意にも時間やコストがかかることから使用環境に適合する言語モデルを使用できることはあまりない.そのため,大規模なデータベースを利用し，書き起こし文に準ずる言語モデルを構築して音声認識を行う研究が行われている．[1][2]

本研究は擬似的な原稿文とフィラーの出現確率を元に言語モデルの作成をして講演音声の認識率向上を目的とするものである．

「日本語話し言葉コーパス」[3]から講演の書き起こし文を抽出し，それを元に半自動で擬似的な原稿文を作成した．講演音声のようにもともと原稿が用意されていたり，発声の内容がある程度決まっていたりする音声においては，例外的な単語の組合せは発生しにくいと考えられる．そのとき，擬似的な原稿文と対象の音声の違いはフィラーや感動詞などの突発的な単語が大半になると思われる．そこで，書き起こし文からフィラーと感動詞の情報を取得する．取得したフィラー・感動詞情報を，擬似的な原稿文を元にした n-gram に反映させて言語モデルを作成する手法を検討し，認識実験を行った．

2. フィラー・感動詞の挿入

前年度でも「日本語話し言葉コーパス」を使用した研究を行っている．[4]書き起こし文と擬似的な原稿文の n-gram を比較して差異を求めた.その差異から近似式を用いて原稿文 n-gram に書き起こし文特有の単語を追加して言語モデルを作成し実験を行った．また，講演音声に対するバックオフ係数が過剰であると考え，その最適値を検討した．2 つの実験から書き起こし文特有の情報の挿入により認識精度が改善されることと，バックオフ係数は従来の約 1/32 倍が最適値になることが確認できた．

この前年度研究では書き起こし文と原稿文の差異である単語すべてを,言語モデルに反映させている．しかし，この 2 つのデータの違いは主にフィラーや感動詞であるが，それ以外にも多くの単語が検出される．これらをすべて訓練データから予測し，言語モデルの修正を行うことが望ましいが，正確な推定をするのは難しい．

これを考慮し，本研究では原稿文 n-ram に挿入するものを種類・頻度などの傾向を特定しやすいフィラーと感動詞の uni-gram のみに限定する．

3. 言語モデル作成

図 1 に言語モデル作成の流れを示す．まず「日本語話し言葉コーパス」の講演データから書き起こし文を自動で作成した．その書き起こし文から言いよどみ，言い直し，感動詞と接続詞の一部，フィラーを半自動で取り除いて擬似的な原稿文を作成した．その書き起こし文と擬似的な原稿文のそれぞれから n-gram データを作成する．書き起こし文 n-gram からフィラー・感動詞の uni-gram の種類・頻度を取得する．取得した uni-gram を原稿文 uni-gram に挿入し言語モデルを作成する．また，挿入する uni-gram の頻度を 1/2 倍，2 倍，4 倍，・・・，64 倍とすることでフィラー・感動詞の影響を変化させ，それぞれの言語モデルを作成する．

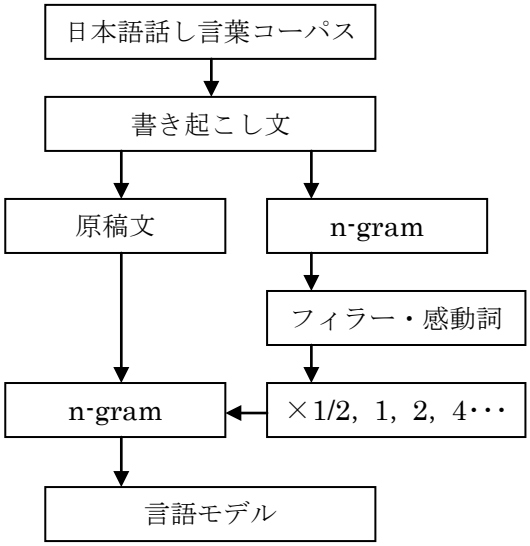

図 1　言語モデル作成

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

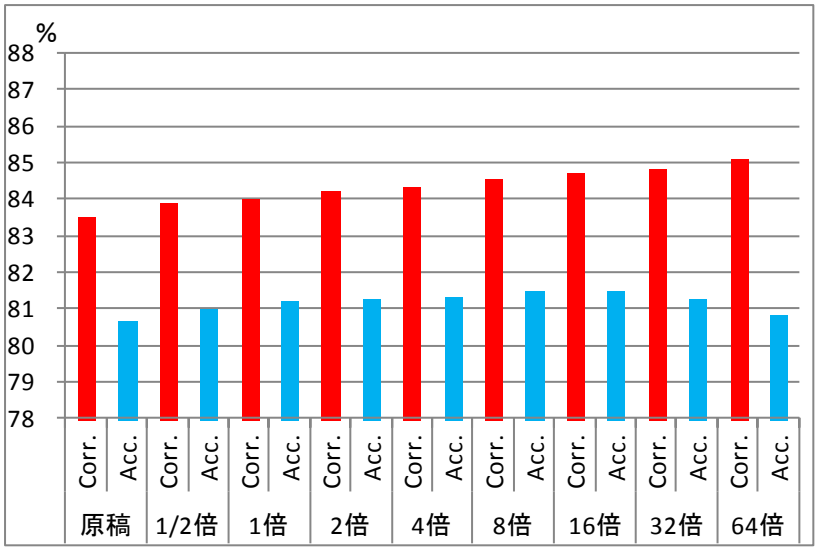

図 2　uni-gram 追加

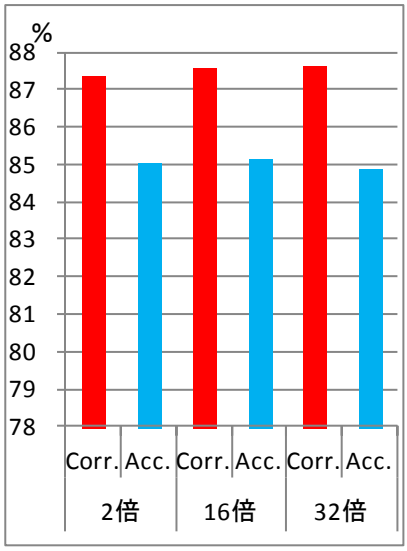

図 3　バックオフ係数 1/32 倍

4. 実験

4.1. 実験方法

本研究では認識器として大語彙連続音声認識エンジン Julius[5]を使用して認識実験を行う．認識対象は「日本語話し言葉コーパスの」の講演音声を用いる．評価方法は単語正解率 Corr.「正解単語数/対象単語数」と単語正解精度 Acc.「(正解単語数－湧出単語数) /対象単語数」を用いる．

実験は擬似的な原稿文から作成した n-gram に書き起こし文から抽出したフィラー・感動詞を追加したもので作成した言語モデルと，追加するフィラー・感動詞の頻度を 2 の累乗倍していったモデル，1/2 倍したモデルで認識実験を行い認識率を求める．

4.2. 実験結果

原稿文 n-gram にフィラー・感動詞の uni-gram 確率を追加し，その確率を 1/2 倍と 2 の累乗倍していったモデルの認識実験結果を図 2 に示す．各モデルはそれぞれ 10 講演分の認識結果を平均したものである．Baseline は原稿文 n-gram に手を加えていない言語モデルでの認識結果を示している．

原稿文のみに比べ，フィラー・感動詞を追加したものはすべて Corr.，Acc.ともに改善されている．しかし，1/2 倍では原稿文のみより改善されてはいるが，その改善率は低い．追加するフィラー・感動詞を 2 の累乗倍していった結果は 16 倍まで Acc.が上がり，32 倍以降は Corr.は精度が改善され続けるものの Acc.はそれまでに比べ下がり始めた．

さらに前年度研究で得られたバックオフ係数効果を確認するため，今回の手法でもバックオフ係数を 1/32 倍にして実験を行った．フィラーと感動詞の uni-gram 確率を 2 倍，16 倍，32 倍にしたモデルに対して認識実験を行った結果が図 3 である．どのモデルでも図 2 のバックオフ係数低減前より認識率が向上している．

5. おわりに

今回行った実験では原稿文 n-gram に挿入したのは純粋にフィラーと感動詞の uni-gram のみである．フィラーと感動詞の挿入のみでも講演音声の認識率が改善されることが分かった．また，フィラーと感動詞の確率を 16 倍程度に増やすことでさらに精度の向上が得られることが分かった．さらに今回の手法に対してもバックオフ係数低減の併用が可能であることが確認できた．今後，フィラー・感動詞を含む bi-gram の利用などにより認識率のさらなる向上を目指したい．

参考文献

[1]秋田，河原：統計的機械翻訳の枠組みに基づく言語モデルの話し言葉スタイルへの変換，情報処理学会研究報告．SLP，音声言語情報処理　2005(127)，109-114，2005-12-21

[2]太田，土屋，中川：フィラー予測モデルに基づく話し言葉言語モデルの構築，情報処理学会誌 Vol.50，No.2 ，477-487(Feb.2009)

[3]菊池，塚原，小町，山田，高橋：日本語話し言葉コーパス，国立国語研究所(2004)

[4]三本木，浦谷：原稿の存在する講演音声の認識率向上，言語処理学会第 17 回年次大会発表論文集，pp.89-90，2011 年(学会発表要旨)

[5]李晃伸：大語彙連続音声認識エンジン Julius ver.4，電子情報通信学会技術報告．SP，音声 107(406)，307-312，2007-12-13

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved