原稿の存在す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

音声認識では使用環境に適合する言語モデルを使用することで高い認識率を有することができる．しかし，実際には言語モデルに必要である書き起こし文を入手できることは稀である．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少ない．そのため，書き起こし文を元にした言語モデル，もしくはそれに準ずる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]

本研究では「日本語話し言葉コーパス」[2]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．その原稿文と書き起こし文を比較して，その差異からフィラーなど発話特有の情報を獲得する．その情報を原稿文を元にした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．

また，講演音声など読み上げる内容がある程度決まっている音声においては例外的な単語の組み合わせは発生しにくいと考えられる．そのため，スムージングは未登録語に対応するために行われるが，その時のバックオフ係数は原稿に沿った発話を前提とするときには過剰な値となっていると思われる．そこでバックオフ係数の最適化を検討したのでその結果について報告する．

2.　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」から書き起こし文を作成した．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，フィラーを取り除いて原稿文を作成した．それぞれのn-gramデータを作成してそれらを比較してその差異を比較した．ディスカウントにはウィッテン・ベル法を用いた．原稿を読む際に話し言葉に追加されると考えられるフィラーや感動詞（一部）をn-gramに付加した．原稿文のn-gramにフィラーなどの話し言葉特有の単語を付加し，超過した確率の分だけ他のn-gramから引くことで全体の確率を整えた．

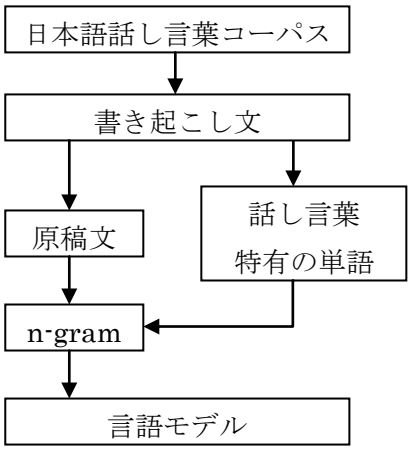

図1　　言語モデル作成の流れ

3.　バックオフ係数の減算

1.で述べたように今回のような音声認識では従来のままのバックオフ係数だと過剰な値になってしまうことが考えられる．図2のようにlog値を−1.0，−1.5，−2.0･･･としていき，本来のバックオフ係数を100%として1/10倍，1/32倍，1/100倍，1/516倍，1/1000倍と減算する．減算した言語モデルをそれぞれ認識実験して認識率にどのような影響があるかを検証した．

答え+コタエ+47　−0.4693

・$\log_{10}X = -0.4693$

$X = 10^{-0.4693} \fallingdotseq \mathbf{0.3393}$

↓

・$\log_{10}X = -0.4693 - 1.0 = -1.4693$

$X = 10^{-1.4693} \fallingdotseq \mathbf{0.0339}$

・$\log_{10}X = -0.4693 - 2.0 = -2.4693$

$X = 10^{-2.4693} \fallingdotseq \mathbf{0.0033}$

図 2　　バックオフ係数の減算

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[3]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」の講演音声を用いる．Julius で認識できる長さに音声ファイルは編集する．評価は単語正解率 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．

実験はフィラー，感動詞などを原稿文から作成した n-gram の uni-gram のみ足したもの，bi-gram と tri-gram にも足したものの二つを行う．さらにそれらのバックオフ係数を 1/10 倍，1/32 倍･･･1/1000 倍としていって認識率を求める．

4.2.　実験結果

表 1 に 10 講演分の認識結果を記す．表 1 の結果は書き起こし文，原稿文の結果と原稿文の n-gram にフィラー，感動詞などの uni-gram のみを加えたものの結果である．Baseline は原稿文を従来のまま言語モデルにしたものを認識した結果である．書き起こし文には及ばないものの uni-gram を加えたものが原稿文のみに比べ改善されたことが確認できた．

表 1　　音声認識実験結果（%）

	書き起こし		原稿		uni-gram	
	Corr.	Acc.	Corr.	Acc.	Corr.	Acc.
1	83.36	77.93	80.97	76.37	81.72	77.46
2	85.93	82.87	81.09	77.82	82.83	80.72
3	82.00	76.86	77.80	72.54	78.44	74.67
4	83.78	78.53	82.63	78.34	83.26	78.49
5	90.19	85.78	88.37	84.64	89.11	86.99
6	89.73	87.62	89.13	87.27	89.66	88.83
7	88.09	85.23	87.06	84.88	87.42	85.74
8	82.32	79.40	73.62	70.19	75.00	72.43
9	87.97	84.47	87.24	83.86	87.81	85.87
10	86.16	82.78	81.37	78.02	82.83	80.46

5.　おわりに

今回行った実験は言語モデルのサイズが非常に小さく検証に不十分だった可能性がある．今後，言語モデルのサイズを大きくし，実験量も増やすことで大語彙連続音声認識として認識率が出せるかを確認したい．

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]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)

[3]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研究報告．SP，音声 107(406)，307-312，2007-12-13

[4]浦谷，小早川：対話システムにおける音声認識の改善を目的としたバックオフ係数の検討，2006 年言語処理学会年次大会 B1-2