論文 / 著書情報
Article / Book Information

論題(和文)	新聞読み上げタスクを用いた大語彙連続音声認識における言語モデルの検討
Title(English)	
著者(和文)	森岳至, 大附克年, 松岡達雄, 古井貞熙, 白井克彦
Authors(English)	SADAOKI FURUI
出典(和文)	日本音響学会 1996年春季講演論文集, Vol. , No. 3-8-7, pp. 159-160
Citation(English)	, Vol. , No. 3-8-7, pp. 159-160
発行日 / Pub. date	1996, 3

3—8—7

新聞読み上げタスクを用いた大語彙連続音声認識における言語モデルの検討*

⊙森岳至[1] 大附克年[2] 松岡達雄[3] 古井貞熙[1,3] 白井克彦[2]
([1]東工大 [2]早大 [3]NTT ヒューマンインタフェース研究所)

1 はじめに

大語彙連続音声認識においては，探索空間を絞り込むための言語モデルによる制約が重要であるが，知識量が膨大であることから人手による作成は困難である．最近，大規模な電子化テキストが利用可能になったことから，これらを用いて統計的言語モデル (N-gram) を推定し，大語彙連続音声認識に適用する研究が，米国 ARPA の WSJ タスク [3] を始めとし，英語，仏語，独語，伊語などに対して行われている [4]．しかし，これまで日本語を対象とした，これに類する大語彙連続音声認識の研究は試みられていない．本稿では，5 年分の新聞記事データベースを用いて unigram,bigram,trigram の言語モデルを推定し，テストセットパープレキシティにより評価した．また，新聞記事読み上げデータベース [6] を用いて，大語彙連続音声認識における言語モデルの効果を評価した．

2 言語モデルの推定

2.1 テキストデータ

本研究では，日本経済新聞の記事 5 年分 [2] を使用し，全体の 95% にあたる前半 57ヵ月分の記事を言語モデル学習用 (training set) とし，残りを評価用 (test set) とした．

日本語の文章は，英語などと異なり単語境界が明白でなく，また単語の定義もあいまいである．ここでは単語として形態素を採用し，テキストデータに対して形態素解析を行ない，単語に分割する．用いた形態素解析ツールの，日経新聞の記事に対する性能を表 1に示す．

また，N-gram を推定する際に不要となる記号等を取り除くために，以下のような前処理をテキストに施す．

- ( ) 【】 [] <> 《》 〈〉 は中身ごと削除

  これらは前の名詞の説明等に使われているだけなので，中身ごと削除しても文としての意味が変わらない

  (例 -1) 西岸以外の整備計画は建設省の都市公園事業や県の単独事業を利用、「平成十年をめどに国際的にも通用するような立派な水きん公園にしていく」（吉川正夫町長）考えだ。
  ⇒ （発言者名を削除）

表 1: 形態素解析ツールの性能

語彙数	約 25 万語
解析正解率	約 95%
解析時間	約 58 時間 (210M 形態素の解析)

- ′ ″ #@☆★○●◎◇◆などは削除

  これらの記号自体は注意を引きつけるためのものであり，読みようがないので，テキストから取り除く．

  (例 -2) 自民党及び無所属の派閥は［三］＝三塚派、［宮］＝宮沢派、［渡］＝渡辺派、
  ⇒ （略称の [] を削除）

- ＝±×÷∞℃＄％＆〒などはそのまま残す．

  これらの記号はそれ自体に意味があり，読む必要がある．

  (例 -3) それぞれ五九・八％の人が好きな理由として挙げている。

このような処理を施した後，形態素解析を行なった．

一文に含まれる形態素数が極端に多い文は，名詞の羅列のような，文構造に意味のない文が多い (人事欄等)．そこで，形態素解析の結果から，一文中に含まれる形態素数の分布を正規分布とみなしたときの，一文あたりの形態素数が平均値 $\pm 2\sigma$(95.5%:1~53 形態素) に含まれる文のみを N-gram の推定に用いた．以上のような前処理を行なった後の学習用テキストに関するデータを表 2に挙げる．

テキストに関して，以下のような問題があった．

- 表記のゆらぎ

  漢字とひらがなに関するゆらぎ (例:「よく - 良く」)
  ふりがなのゆらぎ (例:「現れる - 現われる」)

  しかし，頻度が等分に割れると，単語の生起確率が小さく計算されてしまうが，そのような例は少なく，統計的に影響はないと考えられる．

- 前処理による問題

  ・「(1)」，「(2)」などを削除することにより，文構造に曖昧性が生じる．

  (例) 日本側が前日に続き（1）年金・投資信託など 資産管理業務（2）社債市場（3）証券会社の越境取引 ——の三項目に関する規制緩和措置を説明した。

  この例では，「(2)」の削除により，「資産管理業務社債市場」を (固有) 名詞とする構造が生じる．

  ・同じ記号について，削除するべき文と削除すべきでない文が存在する．

*A Study of Language Modeling for Large-Vocabulary Continuous Speech Recognition Using A Read-Speech Corpus.
By Takeshi Mori[1],Katsutoshi Ohtsuki[2],Tatsuo Matsuoka[3] ,Sadaoki Furui[1,3]and Katsuhiko Shirai[2]
([1]Tokyo Institute of Technology,[2]Waseda University, [3]NTT Human Interface Laboratories)

表 2: 前処理後のテキスト

文章数	6.8M 文
形態素数	180M 形態素
ファイルサイズ	660MBytes

(例-1) 〈十五カ月予算〉今年度第三次補正予算と来年度 当初予算を合わせて九四年一月から九五年三月までの 十五カ月間を対象にした予算編成を目指す。

(例-2) 主な著書に「〈ポスト個性化〉の時代一高度 消費文化のゆくえ」「アイドル工学」など。

例-1での「<>」括弧は中身ごと削除してもよいが，例-2では括弧内の削除により文構造が壊れる．

このような例は非常に少ないので，統計的に影響はないと考えられる．

2.2 N-gram の推定結果

N-gram は単語を出現頻度順にソートした単語頻度リストの上位 7K 語 (カバー率:90.3%)，30K 語 (カバー率:97.5%) を語彙サイズとして推定した [6]．結果を表 3と表 4に示す．なお，trigram と bigram に対しては，Katz の backoff-smoothing [5] を行なっている．

trigram は bigram に比べ，総種類数が約 6~8 倍とかなり多く，詳細なモデルを作成することが期待できる．半面，平均頻度が約 5~7 回と，一モデル当たりの学習量が少ないモデルが多く，信頼性は低下する．

WSJ タスクでは，区読点や引用符を読み上げる形式 (Verbalized Punctuation:VP) と，読み上げない形式 (Non-Verbalized Punctuation:NVP) の 2 通りの言語モデル，評価用テキストを使用している．

本研究においては，N-gram を推定する際には句読点を含めた学習を行なっているので，WSJ タスクとの比較には句読点を考慮した場合 (VP) と比較するのが妥当である．しかし，言語が異なる上に単語の定義が異なるため，厳密な比較を行なうことはできない．

3 連続音声認識実験

今回，新聞記事読み上げデータベースの 7K セット 10 名分 200 発話 (open,close 各 100 テキスト) について連続音声認識実験を行なった．ここでは言語モデルとして，bigram を使用した．音響モデルは [6] における音素認識実験で最も性能の高かった文脈依存 HMM(モデル数 =748) を使用した．また，特徴量としては，16 次元の LPC ケプストラムおよび Δ ケプストラムを使用した．実験結果を表 5に示す．

bigram を言語モデルとして使用することにより，文法無し (NG) の場合に比べ，約 67% 誤り率が改善し，非常に大きな効果があることが示された．

4 まとめ

本稿では，日本語大語彙連続音声認識における言語モデルを大規模なテキストコーパスから推定し，新聞読み上げデータベースに適用した認識結果について報告した．

今後の課題としては，trigram や CFG などの言語モデルによる，より広範囲な単語間の制約についての検討などが考えられる．

謝辞

形態素解析ツールを提供していただいた NTT ヒューマンインタフェース研究所映像処理研究部の田中一男主幹研究員に感謝します．テキストデータの使用を許諾していただいた日本経済新聞社に感謝します．また，日頃御指導いただく NTT ヒューマンインタフェース研究所古井特別研究室，東工大古井研究室の皆様に感謝します．

参考文献

[1] 大附, 森, 松岡, 古井, 白井, "新聞記事を用いた大語彙連続音声認識の検討", 信学技報, SP95-90.

[2] "日本経済新聞 CD-ROM 版 1990 年版～ 1994 年版", 日本経済新聞社, 1994-1995.

[3] "The Design for the Wall Street Journal-based CSR Corpus", D.B.Paul,J.M.Baker, Proc. ICSLP '92, 1992.

[4] "Speech Recognition of European Languages", Lori Lamel,Renato De Mori, Proc. IEEE ASR Workshop, 1995.

[5] S.M.Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer", Trans. ASSP-35, 1987.

[6] 大附, 森, 松岡, 古井, 白井, "新聞記事読み上げタスクを用いた大語彙連続音声認識における音響モデルの検討", 音講論, 1996-3.

表 3: 推定した N-gram のデータ

size	lang. model	総種類数	平均頻度
7K	unigram	7K	24338.0
	bigram	2.1M	65.1
	trigram	17.1M	7.2
30K	unigram	30K	6121.9
	bigram	4.9M	33.8
	trigram	30.5M	5.1

表 4: テストセットパープレキシティ

日経			WSJ			
size	lang. model	test-set perp.	size	lang. model	test-set perp. VP	test-set perp. NVP
7K	UG	597	5K	UG	-	-
	BG	82		BG	80	118
	TG	38		TG	44	68
30K	UG	693	20K	UG	-	-
	BG	124		BG	158	236
	TG	64		TG	101	155

UG:unigram,BG:bigram,TG:trigram

表 5: 認識結果

言語モデル	training set		test set	
	Cor.	Acc.	Cor.	Acc.
NG	24.5	23.0	23.1	22.1
BG	77.2	73.5	76.0	72.5

Cor.= 正解率 (%Correct), Acc.= 正解精度 (Accuracy)