確率的言語モデルによる自由発話認識に関する研究

鳥取大学　工学部　村上仁一

1997/2/27

研究の背景および目的

文法的な言語モデル
ネットワーク文法
文脈自由文法
文脈依存文法

問題点：ルールの維持管理
（人間に大きな負荷が必要）

確率的な言語モデル
N-gram (bigram,trigram)
確率付きネットワーク文法
確率付き文脈自由文法
確率付き文脈依存文法

問題点：確率の付与方法
（基本的には大量の
テキストが必要）

最近の傾向

コンピュータの発達：

CPU、メモリー、DISKのコスト小

大量のテキストデータベース：

英語：Brown corpus,AP corpus

日本語：なし → 新聞記事のコンピュータ化

CD-ROM販売など

↓

確率的言語モデルの有効性の定量的な評価

研究の位置づけ

・確率的言語モデルの有効性の定量的な評価
  ・日本語のN-gram
  ・日本語の確率付ネットワーク文法
  ・音声認識への応用（自由発話への適用）
・自由発話の調査
  ・音響的な特徴
  ・言語的な特徴
  ・アクセントの情報量など

同類の研究（英語では古い）
  ・N-gramモデル　シャノンの情報量
  ・N-gramの音声認識への適用（IBM）
    1970年代　ただし発表は1980

年代後半
  ・N-gramの形態素解析への適用
    ATT Ken-church　1980年代後半

結論

2. 連続音声認識システムに使用するアルゴリズム

- HMMの説明
- Baum-Welch アルゴリズム
- Vietbi アルゴリズム
- One-pass DP
- その他、高速化の手法

3. 日本語のN-gramによるモデル化

- データ量に対するエントロピーの変化
- 新聞記事
- X線CTレポート
- ATRの対話データ

「学習データが増加した場合、全体に占める割合は少ないが、
たえず新しい種類の連鎖が出現する」

「言語モデルとしてのマルコフモデルの妥当性」
- 滅多に出現しない言語現象は、あえてモデルに適合させる必要がない

結論

4．N-gram を用いた音声認識

・かな、漢字、品詞のtrigramの有効性
(新聞記事・シミュレーション)

・X線CTにおけるbigramの有効性

・ATRの国際会議の申し込み文の入力におけるtrigramの有効性
「音声認識においてtrigramは有効」

5．自由発話の音声認識

・言語モデル(Perplexityの低い言語モデル)

・言い淀み、言い直しの対処
Garbage model 音素スキップ
「単語のtrigramは有効」
「音素スキップが有効」

結論

6．自由発話音声における音響的･言語的な特徴

･音響的
　発話速度
　音素認識率
　「自由発話は朗読発声とあまり差がない」

･言語的
　間投詞の出現率　　「文の約40%に出現」
　言い誤りの出現率　　「文の約10%に出現」

7．音声におけるアクセント情報の持つ情報量の考察

･漢字かな変換を用いた測定方法
　「韻律は多くの情報量を持つ」

8．Ergodic HMMを用いた未知･複数信号源クラスタリング問題の検討

･複数話者が話されたときの話者認識
･Erogodic HMMと、その利用
　「話者特徴量の抽出に長時間分析が有効」
　「高い尤度を持つ初期モデルの選択が有効」

HMM　(Hidden Markov Model)

・状態を陽にしない状態遷移オートマトン

S1,S2,S3 :状態
a,b: シンボル出力確率

Baum-Welch学習

シンボル系列が与えらた時、尤度を最大にするように
パラメータ(初期状態確率、状態遷移確率、シンボル出力確率)
を学習

- Forwardアルゴリズム
  前から入力した時の尤度
- backwardアルゴリズム
  後ろから入力した時の尤度
- Forward-Backwardアルゴリズム

連続音声認識のアルゴリズム

(与えられたデータの尤度が最大になる状態系列を推定)

- Tree-- Trellis サーチ (フルサーチ)
- Viterbiサーチ(One Pass DP)
- (Level building)

認識アルゴリズムの改良点

- N-bestサーチ
- 経路計算
- beam-search
- ビームの枝刈りの方法
- 近接したフレームにおける言語モデルの
  確率値の再計算
- trigramの値のindex方法(完全hash)
- log計算
- 認識単位(音素)
- 遅延言語処理

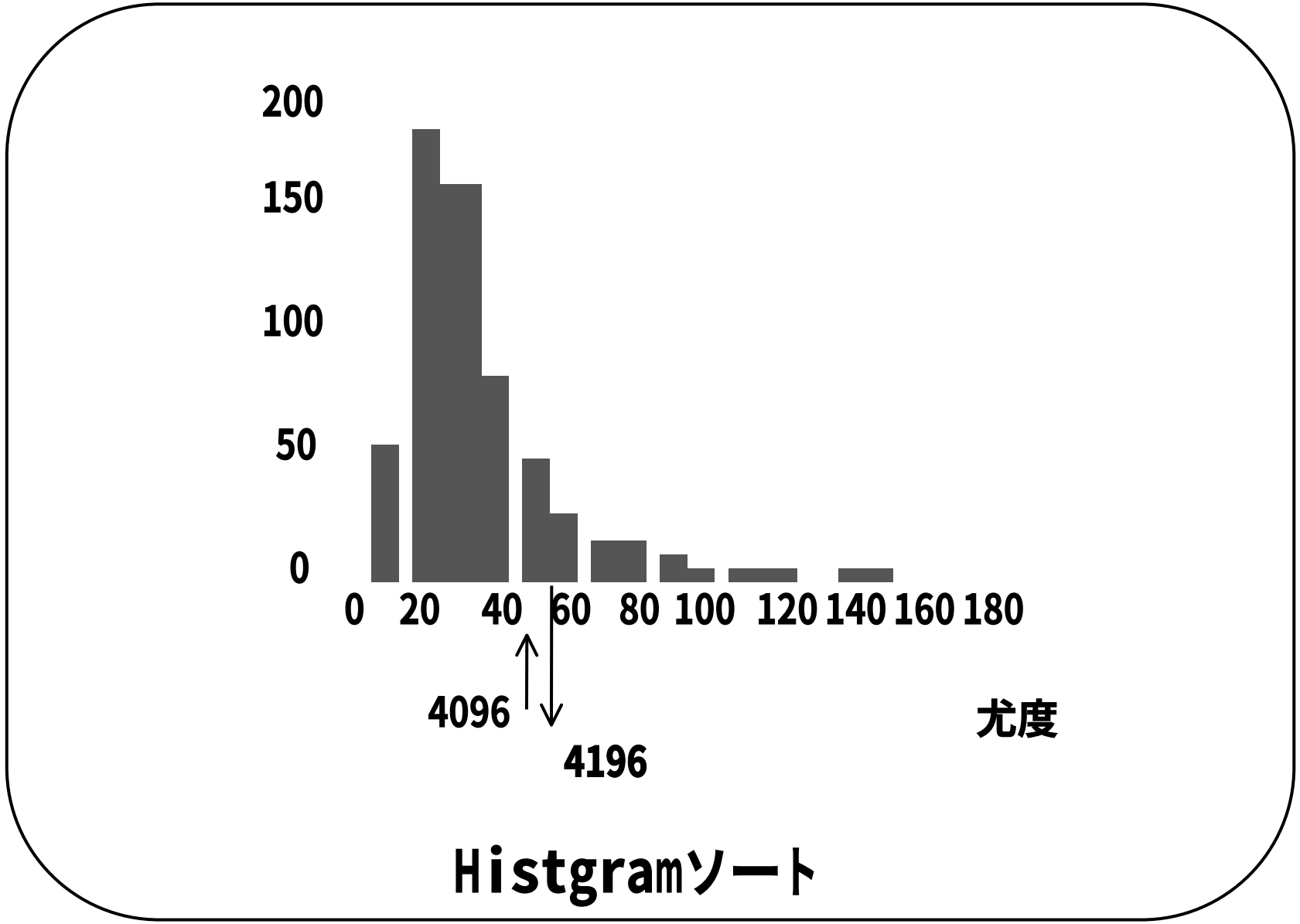
ビームの枝刈りの方法

1) 尤度　　予め決めておいた値で計算を打ち切る
利点：計算速度　　欠点：不安定

2) 幅　　一定の幅で計算を打ち切る
欠点：フレームごとのソーティングが必要

改良点)　　着目点：正確なビーム幅は不要
ビーム幅のスレッシュホールドをhistgramで
計算して枝刈り

→　計算量が通常のフルソーティングと比較して大幅に減少



Histgramソート

改良点

- N-best tサーチ
- 経路計算
- Beam-search
- ビームの枝刈りの方法
  - 尤度の値で枝刈り
  - ビーム幅で枝刈り
- 近接したフレームにおける言語モデルの値の再計算
- trigramの値のindex方法(完全hash)
- log計算
- 認識単位(音素)
- Look ahead処理(言語モデルを後で計算)

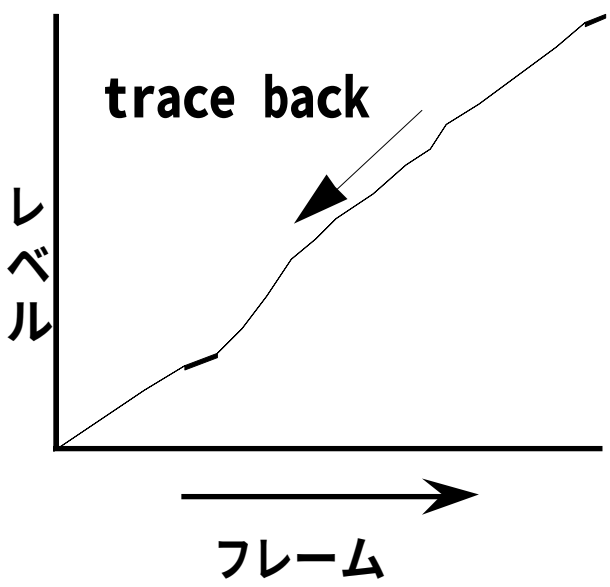
ビーム幅の決め方

1) 尤度　予め決めておいた値で計算を打ち切る
　利点：計算速度
　欠点：不安定

2) 幅　一定の幅で計算を打ち切る
　欠点：フレームごとのソーティングが必要
　　→ これが大きなオーバーヘッド？
　改良点：ビーム幅のスレッシュホールドを計算して枝刈り

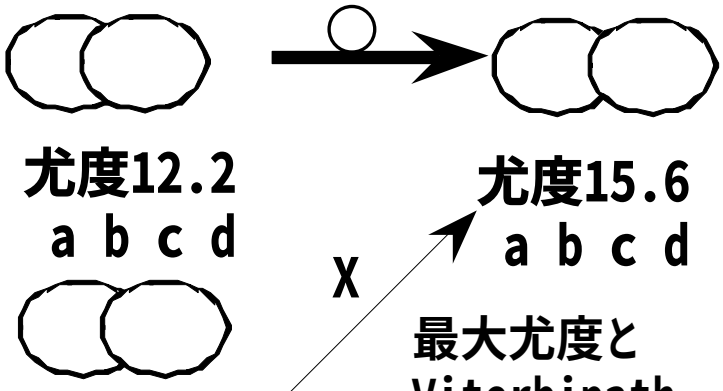
$O(\log_2 N)$

Viterbiサーチの経路計算　(メモリー量の削減)



最大尤度の計算を終了後
トレースバック

従来法



各フレーム･各状態においてトレースバックなしで最尤の単語列がわかる

改良法

結論

9．Ergodic HMMを用いた確率付きネットワーク文法の自動獲得

・確率付ネットワーク文法とHMMの類似性

・入力　品詞　｜　単語

・状態数が大きいときのBaum Welch学習

Baum-Welch学習により文法の自動獲得が可能」

「獲得された言語モデルは音声認識に有効」

3．日本語のN-gramによるモデル化

- データ量に対するエントロピーおよびカバー率の変化
- 新聞記事
- X線CTレポート
- ATRの対話データ

学習データ量に対するEntropyの変化

entropy		
Unigram Σ i	p(wi)	log p(wi)
Bigram Σ(i,j)	p(wi,wj)	log p(wj | wi)
Trigram Σ(i,j,k)	p(wi,wj,wk)	log p(wl | wi,wj)
4-gram Σ(i,j,k,l)	p(wi,wj,wk,wl)	log p(wl | wi, wj,wk)

カバー率

カバー率Ｘ％の種類の数

入力データのＸ％をカバーするのに必要最小限の
マルコフ連鎖の種類の数

例:unigram　　データ(a b a b a a c d)

カバー率　　100%の種類の数4 (a b c d)
カバー率　　75%の種類の数2 (a b)
カバー率　　50%の種類の数1 (a)

新聞記事　　(７４日分)

・コンピュータによる形態素解析
・コンピュータによる品詞および読み (音節) の付加
・文節単位

全データ量約１７０万文字 (漢字かな)

音節

長音・促音を１音節と計算
記号・外国語読み・数詞は削除
種類の和　　１１１音節

漢字仮名

漢字JIS１級、約３０００文字

品詞

人名・地名などの意味的なカテゴリを含んで約４５０種類

X線CT所見作成の文章

Total 約25万文字
人手による文節区切り

音節

“mass effect”,“large magna”などの外来語が数多く出現
音節の種類の数は118音節

漢字かな

外来語はアルファベットの全角文字に変換
(例 “mass effect” は “MASS EFFECT”)

単語

語彙数 約3000
ただし、出現率が高い 100文節は単語として登録
(例 “脳実質を”)

X線CT所見作成の文章

頭部ＣＴ　　単純および造影

1、3月13日のＣＴと比較した。

2、スライスのレベルが若干異なっているので正確な比較はできないが、鞍上槽の正中からやや右上方へ向かって進展している増強効果を示す腫瘤の大きさは本質的に変わっていない。ただし前回のＣＴでこの結節性腫瘤の右前方に見られた嚢胞性の成分については今回は描出されていない。

3、側脳室の大きさ形も前回と同様である。

impression.....

鞍上槽の頭蓋咽頭腫の残存については明らかな変化はないが、右後方に見られた嚢胞性成分が消失しているかもしれない。

ATRの言語データベース

発話内容

1　国際会議の申し込みに関する参加者と事務局の対話
（使用データ：語彙6000）
2　旅行に関する旅行会社と客の対話
3　ホテルの予約に関するホテル従業員と客の対話
4　学校訪問の打ち合せに関する学生同士の対話
5　ＮＨＫラジオ第一放送の番組「ふるさと通信」における
アナウンサと全国各地の通信員などとの電話対話

発話環境

1　通常の部屋　　大部分が家庭用のカセットテープレコーダで録音。
外来雑音も混在。
2　スタジオ録音　ＤＡＴで録音。非常に明瞭。
（遮音室）

ATRの言語データベース

国際会議の申し込みに関する参加者と事務局の対話
総語彙6000総単語数約20万

・[あっ、あえーっと] そちら第1回の通訳電話国際会議の事務局でしょうか。
・はいそうです。
・[えーっとちょっと] その会議のことでねあのー登録のことでお伺いしたいんですが。
・はい。
・どうぞ。
・[えーっと] 今手元にあの登録用紙があるんですけれども
・[えーっと] その中でちょっとあのクレジットカードをね
・[あのー] クレジットカードの名前となんかナンバーを書くところがあるんですが
・はいそうです。
・[えーっと] それをちょっとクレジットカードを持っていない者がいるんですけれどもその場合はどうなんでしょうか。
・はい。

X線CT所見における文節音声認識の実験

目的：単語bigramの有効性

入力文　　X線ＣＴ所見作成の文章
基本構成　単語ＨＭＭ＋単語bigram
　　＋Viterbi search (one-pass DP)
語彙数　　約3000
特定話者
文節発声

bigramの学習単語数　約25万単語

	unigram	bigram	trigram
音節	5.5	3.21	1.57
漢字かな	7.3	2.55	1.61
単語	8.13	3.75	2.61

X線CT所見作成 (エントロピ)

	unigram	bigram	trigram	4-gram
音節	5.67	4.29	2.94	2.16
漢字かな	8.15	4.45	2.87	2.29
品詞	5.57	2.69	2.03	1.63

新聞記事 (エントロピ)

新聞記事とX線CT所見作成の比較
X線CT所見作成の文章は新聞記事と比較して単純

入力データ量に対するマルコフ連鎖確率値の変化のまとめ

1　エントロピーとカバー率

安定になるまでの学習データ数

エントロピー　<カバー率>

エントロピーだけでなく、カバー率も調査する必要あり

2　カバー率100%と98%

学習データが増加した場合、全体に占める割合は少ないが、
たえず新しい種類の連鎖が出現する

言語モデルとしてのマルコフモデルの妥当性

・滅多に出現しない言語現象は、
あえてモデルに適合させる必要がない

入力データ量に対するマルコフ連鎖確率値の変化のまとめ

3　新聞記事とX線CT所見作成の比較
X線CTの所見作成の文章は新聞記事と比較して文章が単純

漢字かな bigram　2.55　<　4.45
(X線CTの所見作成)　(新聞記事)

4　形態素解析プログラムの精度
形態素解析プログラムの精度　単語認定率で約95%
人手によって文節単位に区切られた時の値との差
品詞のtrigramに有意性が見られない？

5　ATRの国際会議における単語trigramの信頼性
低い(データ量が必要)

word trigram

$P(W_1,W_2,W_3,W_4,..,W_n) = \prod (W_i \mid W_{i-2}\ W_{i-1})$

利点

Viterbiサーチ(One-Pass DP)と親和性が高い。

単純、学習が容易

評価関数

$$\Sigma \log(P(w)) + \alpha \Sigma \log(P(W_i \mid W_{i-2}\ W_{i-1}))$$

↑ 音響尤度　↑ 結合値　　↑ 言語の連鎖確率

4. N-gramを用いた音声認識

・新聞記事の入力における　かな、漢字、品詞のBigram, trigramの有効性（シュミレーション）
・X線CTにおけるbigramの有効性
・ATRの国際会議の申し込み文の入力におけるtrigramの有効性

評価関数

$$\Sigma \log(P(w)) + \alpha \Sigma \log(P(W_i \mid W_{i-2}\ W_{i-1}))$$

↑ 音響尤度　↑ 結合値　↑ 言語の連鎖確率

学習データ量に対するマルコフ連鎖確率値の変化
(新聞記事)

・新聞記事の入力
　　単音節　(文節単位)
　　１０位以内に正解の音素がある。

・コンピュータによる形態素解析
・コンピュータによる品詞および読み (音節) 付加
・全データ量　約１７０万文字 (漢字かな)

まとめ

- 仮名、漢字、品詞のtgrigram →有効
- 音節選出型と直接選出型では差が小。
- 音節のtrigram closed data 79%(文節認識率)
  open data 56%
- 漢字のtrigram closed data 83%
  open data 65%
  (音節のtrigramより有効)
- 品詞のtrigram
  open dataとclosed dataの差が小
  有効性 小

文(文節)音声認識

音響処理だけでは認識性能は低い

↓

言語処理による認識性能の改善

↓

単語のbigram,trigramが有効

↓

bigram,trigramは統計量

↓

どれだけの学習データ量が必要？

認識実験結果　　文節認識率(%)

学習単語数3　　HMMとBIGRAMの結合値32

duration controlあり　　2 best search

		1位	4位
text-closed data	正常所見	100.0%	(100.0%)
	異常所見	84.2%	(89.5%)
text-open data	正常所見	94.6%	(97.3%)
	異常所見	77.0%	(86.9%)

認識実験結果 3　　文節認識率 (%)

duration controlあり　　8 best search

		1位	4位
text-closed data	正常所見	100.0%	(100.0%)
	異常所見	78.9%	(89.5%)
text-open data	正常所見	89.2%	(94.6%)
	異常所見	%	(%)

認識実験結果 2　　文節認識率(%)

duration controlあり　2 best search

		1位	4位
text-closed data	正常所見	82.6%	(95.7%)
	異常所見	76.3%	(86.8%)
text-open data	正常所見	86.5%	(89.2%)
	異常所見	68.9%	(77.0%)

まとめ

学習データ量に対するマルコフ連鎖確率値の変化
X線CT所見作成の文章
マルコフモデルの連鎖確率値の信頼性を調べるためにはエントロピーだけでなく、頻度別出現率も調査する必要がある

単語のHMMと単語のbigramを用いた文節音声認識システム
認識単位を単語とした場合、良好な文節認識性能が得られる

HMMの学習用の音声データが1つでも、認識はある程度可能

Word Trigramの研究

1回のForward探索(Viterbiサーチ)で
word trigramを用いた連続大語彙音声認識系は
構築されていない.

主な代行手段

IBM	孤立単語発声(1986)
鹿野	カテゴリを用いたword trigram (1988)
BBN	word bigram + word trigram (1993)

理由の1つ
計算量およびメモリー量が爆発的に増大

ここでの報告

1　HMMとtrigramを組み合わせた文音声認識実験

2　　ポーズの処理　　・ポーズのスキップ
　　・ポーズの学習

文音声認識の実験条件

基本アルゴリズム	Continuous mixture HMM+Beam search +word trigram
Mixture数	最大14 (各音素によって変化)
1音素あたりの状態数	4-state3-loop left-right model
使用パラメータ	LPCケプストラム16次 + パワー +デルタパワー+デルタケプストラム16次
フレーム間隔	10ms
フレーム周期	5ms
HMMの学習音声 (特定話者)	単語発声(5240単語)
(不特定話)	単語発声(12名736単語)
音素カテゴリ数	52音素
認識単語数	1567
ビーム幅	4096
言語情報	単語のtrigram
実験文数	261文
発声様式	朗読発話
発声内容	国際会議の申し込み (通称モデル会話)
trigramの計算に使用したデータ	ATRの対話データベース (国際会議の申し込み)
フロアリング	171978単語 exp(-1000.0)

ポーズ処理 1　　ポーズのスキップ

文中に存在するポーズ.

例　　住所は　　pause　大阪市　　pause

電話番号は　　pause　　３３９の　　pause

対策

音響処理では，ポーズを認識

言語処理ではポーズを無視するように計算

例　東京都　港区　新橋　　pause　　１丁目

Ｐ（新橋 | 東京都 港区）X 1.0 X Ｐ（一丁目 | 港区 新橋）

ポーズ処理 2　　ポーズの学習

音声のテストデータの先頭のポーズを利用して

ポーズのHMMを学習

まとめ

1　単語のtrigramをもちいた連続音声認識アルゴリズムの報告
　　改良点　　ビームサーチとViterbi Pathの計算方法

2　ポーズの処理（ポーズのスキップおよび学習）
　朗読発話　文認識率　83.9%
　(不特定話者認識,text- closed)

3　自由発話への適用
　冗長語の処理（冗長語のスキップ、音素のスキップ）
　自由発話　文認識率　42.0%
　(不特定話者認識,text-open)

考察

自由発話音声の認識とtext-open data

自由発話：もしtext-closed dataならば、ある程度認識可能

現実：　　text-open dataになる。

今後の研究：

text-closed dataとtext-open dataの認識率の差を減少

方法　　大量のテキストデータを収集？

テストデータへの動的な適用？

単語間の距離の測定？

まとめ

1　単語のtrigramをもちいた連続音声認識

2　ポーズの処理（ポーズのスキップおよび学習）
　朗読発話　文認識率　83.9%
　　（不特定話者認識，text-closed）

今後の研究：
　text-closed dataとtext-open dataの認識率の差を減少
　方法　　大量のテキストデータを収集

５　自由発話の音声認識

間投詞「あのー」「えーと」言い淀みや言い誤りおよび言い直し

１）言語モデル認識精度の高い音響モデルを作成することは困難？

→　perplexityの低い言語モデル（単語trigram ）

２）間投詞や言い直しの対応方法

A)　garbageモデル　　B)　音素スキップ

garbageモデル（音響モデルによる対策）

認識アルゴリズム

garbageモデル=1単語

単語のtrigramの連鎖確率値

garbageモデルをスキップ

実験結果のまとめ

１）音素スキップ、garbageモデル、共に有効

２）認識性能　音素スキップ ＞ garbageモデル

３）認識性能　平滑化しない場合 ＞ 平滑化する場合

４）自由発話において　47.7%の文認識率
単語のtrigramの連鎖確率値の平滑化しない場合
音素スキップ

５）意味的に正しいとみなされる文を正解に含めた場合
1位文理解率で約75%、
8位までの累積文理解率は90%

音素スキップ（言語モデルによる対策）

間投詞や言い直し　→　音素系列
言語モデル　→　音素系列をスキップ
（ペナルティ　→　音素のtrigram）

「”東京都“ ”港区“ ”新橋“ ”あのう(anou)“ “一丁目”」
「あのう」→間投詞

言語モデルの連鎖確率値

P(``新橋''| ``東京都'', ``港区'') x P(/a/|/sh/, /i/) x P (/n/|/i/,/a/) x
P(/o/|/a/,/n/) x P(/u/|/n/,/o/) x P(``1丁目'' | ``港区'' ``新橋'')

P(/a/|/sh/, /i/) ＝　ペナルティ、
P(”1丁目” | ”港区” ”新橋”) ＝音素スキップ

自由発話の音声データ

1　朗読発話

テキストを読みあげた音声データ。
間投詞や言い淀み･言い直しなし。
言語モデルに対して　text-closed　データ

2　疑似自由発話

間投詞を含むテキストを読みあげた音声データ。
間投詞を除いて、「朗読発話」と発話内容は同一。
言い淀み･言い直しは無い。

3　自由発話

話者はテキストを覚えて、その意図を理解してから自由に発話した音声データ。
間投詞や言い直しや未知語を含む。
言語モデルに対してtext-openのデータ？
（テキストを覚えて発話したデータであるため、
発話内容はtext-closed データに近い。）

考察　　自由発話の音声認識

全ての音素を完全に認識する必要性なし

意味的に合っている文章を出力自由発話の認識のための言語モデル、

非文を生成しないこと、

perplexityが低いことモデルがカバーできない範囲

→　　garbageモデルや音素スキップ

考察

実験結果

音素モデルのスキップ　＞　garbageモデル

広いビーム幅が必要語彙数が多い場合や
ビーム幅が小さい場合

garbageモデルの　＜　音素モデルのスキップ
となる可能性

まとめ

1　連続音声認識アルゴリズム

Word trigram + Viterbiアルゴリズム

間投詞、言い直しの対策

・garbage model

・音素モデルによるスキップ

4　自由発話の認識実験

ビーム幅16384語彙数435において47.7%の文認識率

意味的に正しい文を正解に含めた場合

1位文理解率で約75%、

メモリ量および計算量を削減した Baum-Welchアルゴリズムの提案と 言語モデルへの適用

1: 目的

状態数が大きいErgodic HMMのBaum-Welch学習

1.2: 問題点

状態数が小さい

→ Perplexityが高い

状態数が多い

→ Baum-Welch学習が不可能

1.3: 解決方法

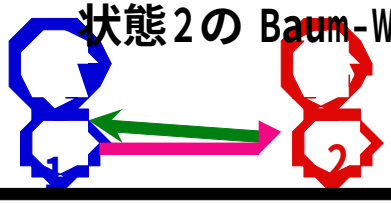
メモリ量および計算量を削減した
Baum-Welchアルゴリズム

小さいシンボル出力確率の削除
(メモリ量, 計算量の削減)

シンボル出力確率が閾値 (10^{-300}) より
小さいとき　0.0
再推定およびメモリから削除。

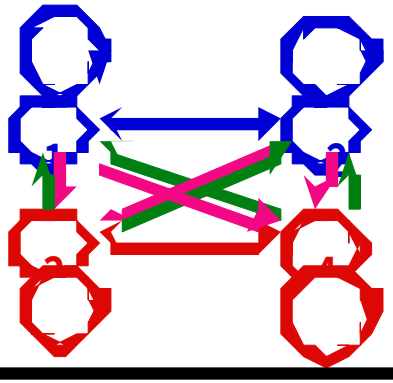
状態数の逐次増加

状態2の Baum-Welch 学習の終了後のパラメータ



$\pi_1^1=0.3$ $a_1^{11}=0.4$ $a_1^{12}=0.6$

$\pi_2^1=0.7$ $a_1^{21}=0.1$ $a_1^{22}=0.9$



状態4の初期モデルのパラメータ

$\pi_4^1=\pi_2^1=0.3$ $a_4^{11}=a_2^{11}=0.4$

$\pi_4^2=\pi_2^1=0.3$ $a_4^{12}=a_2^{11}=0.4$

$\pi_4^3=\pi_2^2=0.7$ $a_4^{13}=a_2^{12}=0.6$

$\pi_4^4=\pi_2^2=0.7$ $a_4^{14}=a_2^{12}=0.6$

ただしシンボル出力確率は乱数を使用

N状態のErgodic HMMのパラメータ

$\pi_N(i)$; i=1,..,N ;初期状態確率

$a_N(i,j)$; i=1,..,N, j=1,..,N ;状態遷移確率

$b_N(i,j,w)$; i=1,..,N, j=1,..,N, w=1,..,V ;シンボル出力確率

V ;語彙数

2N状態のErgodic HMMの

初期状態確率および状態遷移確率の初期パラメータ

$\pi_{2N}(i) = 0.5 \times \pi_N(i/2)$; i=1,..,2N

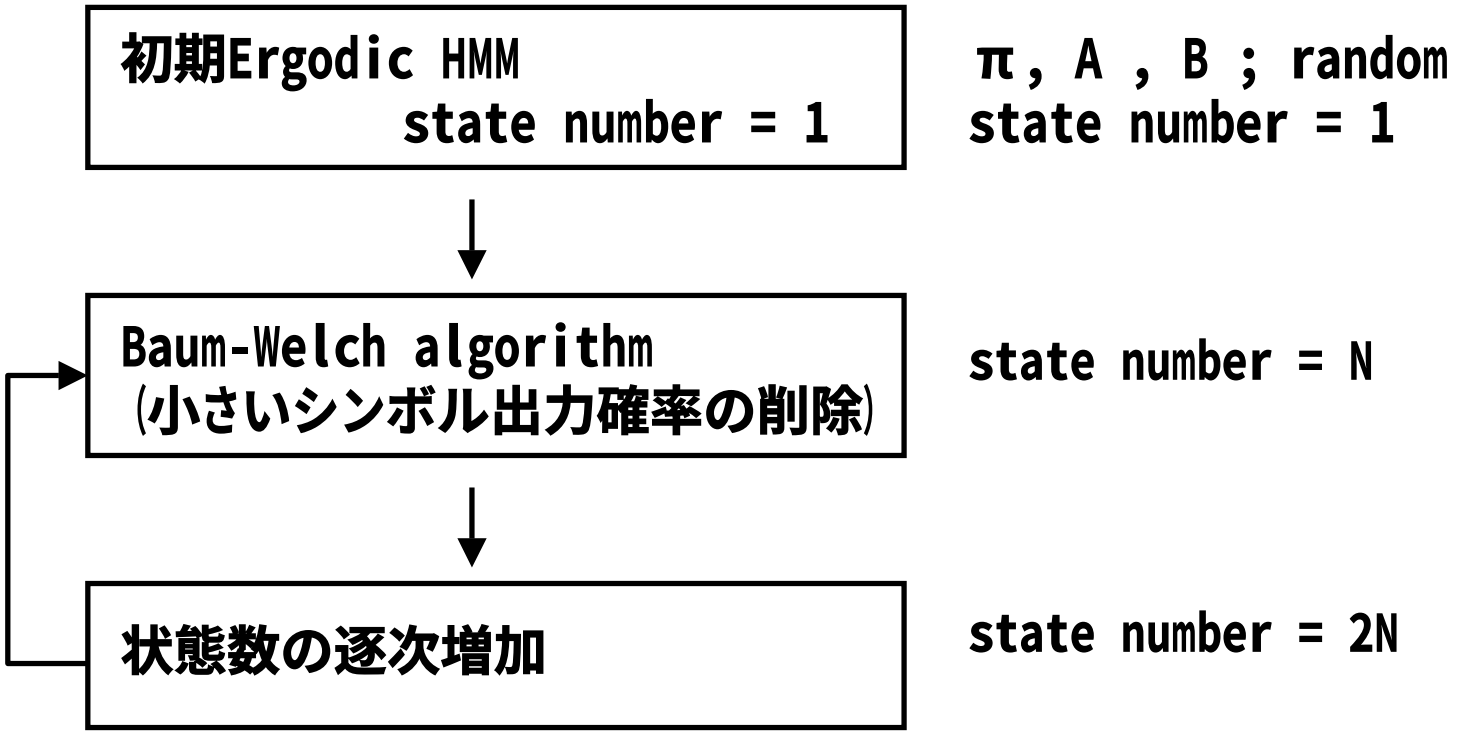
$a_{2N}(i,j) = 0.5 \times a_N(i/2,j/2)$; i=1,..,2N, j=1,..,2N

$b_{2N}(i,j,w) = b_N(i/2,j/2,w) \times random(i,j,w)$;

i =1,..,2N, j=1,..,2N, w=1,..,V. ただし$\Sigma_w\, b_{2N}(i,j,w)=1.0$

" / "切り上げを意味

具体的なアルゴリズム


初期Ergodic HMM
state number = 1
π, A , B ; random
state number = 1
Baum-Welch algorithm
(小さいシンボル出力確率の削除)
state number = N
状態数の逐次増加
state number = 2N

確率付きネットワーク文法の獲得の実験

HMMの構造	状態遷移出力型
学習語彙数	6420単語
学習データ数	8475文
総単語数	57354
HMMパラメータの再学習における Baum-Welchアルゴリズムの終了条件	40回の繰り返し

言語モデル生成実験の条件

まとめ

メモリ量および計算量を削減した
　Baum - Welchアルゴリズムを提案

確率つきネットワーク文法の獲得
　perplexity:
　bigram > Ergodic HMM > trigram

言語モデルとして
　連続音声認識に利用

認識率
　text-closed: bigram<Ergodic HMM<trigram
　text-open: trigram,Ergodic HMM

Ergodic-HMMを用いた確率付き ネットワーク文法の自動獲得

・確率付きネットワーク文法

等値
(シンボル出力確率＝単語の出力確率)

・Ergodic HMM

まとめ

- ４状態　　体言と用言のグループ化
- ８状態　　活用系でのグループ化
- １６状態　　品詞でのグループ化

- 確率付きネットワーク文法の自動獲得が可能

今後の課題

・大量のテキストデータベースの収集

・大量のデータベースを使用したときの

　text-open data とtext-closed dataの差

・N-gramと確率付きネットワーク文法の違い

・未知語処理の処理

・少量のデータベースしかないときの対応

・言語モデルの適応

・日本語における単語の意味

実際のメモリ量および計算量

語彙数１５００：

- 実行空間15Mbyte
- 平均文認識時間１分から２分（HP735）

計算量の考察：

- left-rightパーザタイプの言語モデルは使用可能
- ビーム幅に大きく依存

まとめ

(入力 : X線CT所見作成、ATRの国際会議の申し込み文)

1 Word HMM + Word Bigramの組み合わせで、
良好な文節認識性能が得られる

2 HMMの学習データが少なくても文節認識は可能

3 単語のtrigramが文認識において有効

4 ポーズ処理の有効性
(ポーズのスキップおよび学習)
朗読発話 文認識率 83.9%
(不特定話者認識, test-closed)

新聞記事の例

大蔵省はことし四月から新銀行法が施行されるのに伴い、在日外銀の営業活動を日本の銀行同様に扱うとの基本方針を決め、これを盛り込んだ政令を二月中にも公布する。おもな内容は(1)企業向け貸し出しに対する大口融資規制を在日外銀にも適用し、五年間の猶予期限を設けるなどの配慮をする(2)利益準備金の積み立てを義務づけ、外銀に対する信頼を高める(3)邦銀の支店を買収することや現地法人化を認めるーーなど。大蔵省はこれによって在日外銀に関する法的根拠が明確になるほか、在日外銀の国内活動がしやすくなり、欧米諸国の間に出始めているわが国の金融制度に対する不満を和らげるのに役立つとみている。(在日外国銀行は「きょうのことば」参照

認識単位　　単語

問題点1

学習データ量　　最低3000単語の発音が必要

→　　Fuzzy VQの使用　(小量の学習データ(1 or 3))

問題点2

コンピュータのメモリー量

→　　1単語につき4状態3ループのモデル

3000×256×4×3＝9Mbyte

Word Trigramを利用した文音声認識

実際のメモリ量および計算量

語彙数１５００：

- 実行空間15Mbyte
- 平均文認識時間１分から２分（HP735）

計算量の考察：

- left-rightパーザタイプの言語モデルは使用可能
- ビーム幅に大きく依存

文節音声認識実験のまとめ

1　Word HMM + Word Bigramの組み合わせで、
　良好な文節認識性能が得られる。

2　HMMの学習データが少なくても文節認識は可能

まとめ　(入力:新聞記事)
(シミュレーション)

- 仮名、漢字、品詞のtrigram　→　有効
- 音節のtrigram　closed data　79%　(文節認識率)
  open　data　56%
- 漢字のtrigram　closed data　83%
  open　data　65%
  (音節のtrigramより有効)
- 品詞のtrigram
  open data と closed dataの差が小
  有効性　小

Word trigram

$$P(W0,W1,W2,W3..,Wn) = \Pi\ P\ (W_i \mid W_{i-2}\ W_{i-1})$$

利点

Viterbi サーチ(one-pass DP)と親和性が高い。

単純。学習が容易。

評価関数

$$\Sigma \log(P\ (w)) + \alpha \Sigma \log(P\ (W_i \mid W_{i-2}\ W_{i-1}))$$

↑ 音響尤度　↑ 結合値　↑ 言語の連鎖確率