修　士　論　文

話し言葉音声認識における
文節境界情報を用いた言語モデルの高度化

Improvement of Statistic Language Model Using
Boundary Information for Spontaneous Speech Recognition

2009年2月4日　提出

指導教員　　広瀬 啓吉 教授

東京大学 大学院 情報理工学系研究科

電子情報学専攻

48-076431

細田聖人

内容梗概

音声認識技術は発展を遂げており, それらを応用した様々なアプリケーションを見掛ける事ができる. 一度に大量の音声を処理する事が要求される大語彙連続音声認識においても, 書き言葉音声認識のタスクにおいては, 高い認識率が示されている. しかし, 少し崩れた発話であったりフィラー等の不規則発話要素が含まれる話し言葉音声認識においては, 従来の方式をそのまま用いることでは, 予測精度が低くなる.

我々は, 人間が認識に用いていると考えられる境界情報を言語モデルに組み入れる事で性能の高度化を図ってきた. すでに書き言葉認識において性能の向上がみられ, 話し言葉の音声認識においても境界の検出法を中心に研究が進みつつある.

本研究では, 今まで取り組まれていなかった言語モデルの利用面から, 境界情報の効果的な利用について考察した. 具体的には, N-gram 言語モデルのバックオフによる影響を考慮し, 使用されるモデルそれぞれ (N=1,2,3) に対してその比率と性能の変化を調査した. その結果, 境界を用いた言語モデルにおいては, 精度の高い高次の N-gram はあまり用いられず, 精度の低い低次の N-gram が多く用いられている事が分かった.

調査結果をもとに, 境界情報が効率的に使われる条件を考察し, さらにスパースネスの解消を目的として, 通常の N-gram と境界情報を利用した N-gram の統合モデルを提案した. 実証実験の結果, いくつかの条件においてパープレキシティの改善がみられ, 特に境界の正解を与えた条件で最も良い結果が得られた.

目次

図目次

表目次

第1章

序論

1.1 本研究の背景

音声認識技術は, その応用アプリケーションとして, 様々な状況で活用されうる可能性を秘めているものである. 例えば, 手でのオペレーションが困難な状況において, 機械に音声により指令を送る事や, また膨大な音声情報から自動的にログを取ることなどが挙げられる.

現在, 書き言葉音声を対象とした場合の認識では, 高い認識率が得られているが, 実用的観点から言えば, 従来の主流であった書き言葉認識から, より自然な話し言葉音声に対する認識技術へのニーズが高まっていると言える. 話言葉音声を対象とした場合には, 省略, 言い直し, フィラー挿入などのいわゆる不規則発話に伴い, 認識率が大きく低下する. 書き言葉に比べ話し言葉では, 良好な言語モデルを得ることが困難なことがこの一因と考えられる.

一方で, 我々は人間が音声の認識に利用すると考えられる発話境界情報に注目して言語モデルを高度化する試みを行ってきた. これは, 境界を跨ぐ場合と跨がない場合の単語遷移の様子の違いに着目したもので, 両者の単語 N-gram を個別に取り扱うものである. この言語モデルを用いることで, 書き言葉, 話し言葉認識に対して基礎的な N-gram の改善が図られてきた. アクセント句境界の有無による品詞傾向の違いを利用し, 言語モデルを構築する手法 [2] や, 大語彙コーパスの単語列により学習ができる文節境界境界尤度を利用した言語モデルを認識に用いる手法 [3] を提案し, その有効性を示した. また話し言葉において, 韻律情報を用いた境界検出によるもの [4] も提案された. しかし, 話し言葉音声認識の先行研究における対象は境界検出技術におけるものが中心で, 言語モデルの利用法に関するさらなる考察が求められている.

1.2 本研究の目的

本研究では, 特に言語モデルの利用という面を中心に考察し, 境界情報を用いた言語モデルの高度化を目指す. まずコーパスサイズ, 境界予測単語／品詞などの条件を変えて予測性能の調査を行う. また, N-gram 確率値が見つからない場合に対しては (N-1)-gram 確率値を用いるバックオフスムージングが行われるが, 境界を跨ぐ／跨がない言語モデルに対して, 実際に使われる N の傾向とそれぞれのパープレキシティを調査し, 比較する. それらの結果を踏まえて, 新たな言語モデルの構成法, 利用法について提案を行う.

1.3 本論文の構成

本論文は全 5 章で構成されている. 2 章では大語彙音声認識の基礎を説明するとともに, 言語モデルに注目しその問題点等を述べ, さらに話し言葉音声認識の特徴についても説明する. 3 章では, 境界情報についての基礎的検討を行い, 本研究の先行研究の流れを紹介する. 4 章では, 特に文節境界を用いた話し言葉音声認識において, 品詞遷移傾向といった基

礎的検討から言語モデルのバックオフ調査まで含めた, 改善のための検討を行う. さらに調査結果を元に新しいモデルの提案を行う. 5 章では, 本論文のまとめと結論を示す.

第2章

音声認識システムと言語モデル

2.1 はじめに

本章では, まず大語彙連続音声認識システムの基礎について説明する. 具体的には, 統計的音声認識の仕組みについて, 認識に用いる特徴量の抽出から音響モデルと言語モデルの利用までを簡単に説明する.

続いて, 統計的言語モデルに関する話題に触れる. 現在主流である N-gram モデルについて仕組みを説明する. また N-gram の欠点とそれに対する取り組みについて紹介する.

最後に, 本研究で主に取り扱う事になる話し言葉認識について, 書き言葉認識との違い, コーパスの特徴などから認識における問題点に注目し, さらに改善に向けた取り組みについて考察する.

2.2 大語彙連続音声認識システム

音声認識は, 入力された音声 x に対し, 最も良く合う言語表現 w を推定することであるといえる. 以下では, 音声認識に使う特徴量の抽出過程, 大語彙連続音声認識の統計的基礎, 認識に使われる音響モデルと言語モデルについて説明する.

2.2.1 音声分析と特徴パラメータの抽出

音声認識を行う場合に最初に問題となるのは, 入力された音声信号からどのように特徴パラメータを抽出し認識に利用するか, ということである. 音声は音韻的特徴と韻律的特徴という 2 つの特徴を持っているが, 現在の音声認識では音韻的特徴, すなわち音声のスペクトル包絡を効率よく表現できる特徴パラメータが用いられている. これは, 音韻的特徴が音声を文字表記と対応付けるために最も基本的な特徴だからである.

自然言語のシンボルを構成要素に分解していくと, 最終的に音素にまで分解される. その音素が文字を構成し, 文字が単語を構成し, 単語が文を構成する.

その音素と対応している音声の特徴を音韻的特徴という. 音素は母音と子音に大分されるが, どちらも主に声道の形状を制御することによって表現される. このような音韻的特徴は, 音声信号のスペクトル包絡に表れる. その様子を図 2.1 に表す.

音声分析は, 音声波形から数十ミリ秒程度のフレームを切り出し, その区間を周波数分析することで行われる.

スペクトル包絡を表す特徴パラメータとして現在最もよく用いられているのがケプストラムである. ケプストラムは, 対数スペクトルの逆フーリエ変換として定義される. これを再び図 2.1 を用いて説明する. 図 2.1 の下のグラフの細線は, 上の波形から切り出されたフレームに対して対数スペクトルを求めたものである. これをみてわかるように, 対数スペクトルはスペクトル包絡成分と微細構造とから構成されていると考えられる. この対数スペクトルを逆フーリエ変換するということは, いわば周波数軸を時間に見立てて再度周波数分析することを意味し, それがケプストラムである. 従って, ケプストラムの低次の項には対数スペクトルのなだらかな変化であるスペクトル包絡情報があらわれ, ケプストラムの

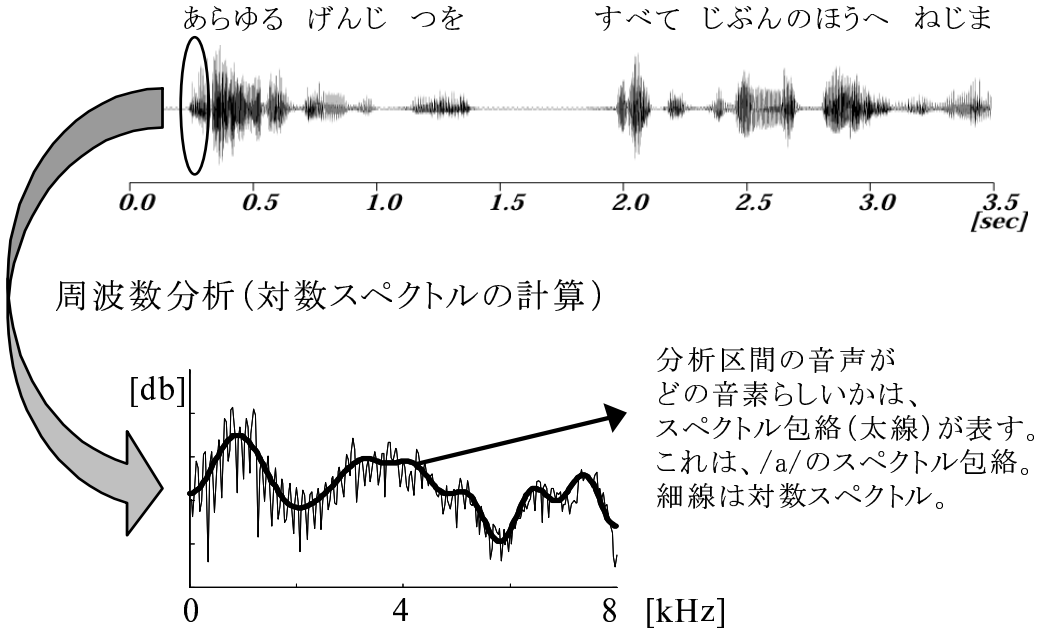

図 2.1: 音韻的特徴とスペクトル

高次の項には細かいスペクトル微細構造が反映されることになる. すなわち, ケプストラム計算によって対数スペクトルの包絡成分と微細構造とを分離することができる. このため, ケプストラムの低次の項を用いることで, スペクトルの包絡情報を効率良く表すことができる.

このようにして抽出されたケプストラムの低次の項（通常 10 次程度までが用いられる）に加えて, 分析区間のパワーも特徴パラメータとするのが一般的である. さらに, 時間と共に変化する音声の動的な性質を考慮して, ケプストラムの時間差分を求めた Δ ケプストラムや, パワーの時間差分を求めた Δ パワーもよく利用される. パワーは韻律的特徴であるが, 基本的には特徴パラメータはケプストラムをベースとした音韻的特徴を表しているといえる.

このようなフレーム単位の分析を, 時間を 10 ミリ秒程度ずつずらしながら入力波形全体に渡って行うことで特徴パラメータの時系列 $X(1), X(2), \ldots, X(t)$ を得ることができ, これが認識に用いられる.

2.2.2　大語彙連続音声認識の統計的手法

音声の特徴量抽出によって, 音素に関する情報を表すケプストラム特徴量が時系列で得られる. このように得られた特徴量の時系列 $X = (x(1), x(2), \cdots, x(t))$ について, 統計的手法を用いて, 最も適合する単語列 $W = (w(1), w(2), \cdots, w(t))$ を推測するのが統計的音声認識の目的である.

すなわち, 観測された X が W である尤度

$$P(W|X) \tag{2.1}$$

を最大化するような単語列

$$\hat{W} = \arg\max_{W} P(W|X) \tag{2.2}$$

を求める問題といえる. W が未知なので式 2.2 の右辺を直接求めることは不可能であるが, ベイズの定理を用いて

$$P(W|X) = \frac{P(X|W)P(W)}{P(X)} \tag{2.3}$$

と変形することで, 次のように定式化することができる.

$$\hat{W} = \arg\max_{W} P(X|W)P(W) \tag{2.4}$$

ここで, 右辺第 1 項の $P(X|W)$ は, ある単語列 W が発声されたときに特徴量の時系列として X が観測される確率を意味する. このような W と X に関する統計モデルを**音響モデル**という. また, 第 2 項の $P(W)$ は言語的性質から求められる統計モデルで, 単語列 W が生成される事前確率を表しており, これを**言語モデル**という. これらの音声認識の一連の流れを図 2.2 に示す.

2.2.3　音響モデル

式 2.4 中の $P(X|Y)$ が与える統計モデルが音響モデルであり, 隠れマルコフモデル（Hidden Markov Model, HMM）によって実現する. これは, ある出力系列が与えられたときに, それを与える隠れ状態を仮定し, その隠れた状態の確率モデルのパラメータを推定することで得られる. 推定されたパラメータと, マルコフモデルの遷移確率によって与えられるモデルが音響モデルである.

ひとつの HMM につきひとつの音素モデルを対応させることで, 音素 w を発声したときに得られる特徴量の系列 $\{x\}$ を得る確率（これが最終的に $P(X|W)$ を与える）を得ることができる. 即ち, $\{x\}$ から w を推定することができる. 通常の音声認識では, X を与える特徴量は, スペクトル包絡に基づいた音韻情報である.

音声認識で最もよく用いられる left-to-right 型の音素 HMM を図 2.3 に示す. HMM は遷移確率 a_{ij} で状態 i から状態 j へ遷移を行い, 状態 i では確率分布 $b_i(X)$ に従って特徴パラメータ X を出力する. HMM のひとつの状態は音声の定常的な部分信号を表し, 状態遷移は信号の変化を表しているといえる. 同じ音素でも状況や環境などによって音声信号は大きく変化するが, 時間的な揺らぎは確率的な状態遷移によって, スペクトル的な変動は確率的なパラメータ出力の分布によってそれぞれ吸収されるので音素の特徴をうまく認識することができる. この様子を図 2.3 に示す. このように, HMM は音響特徴量の生成モデルとして優れた性質を持っている.

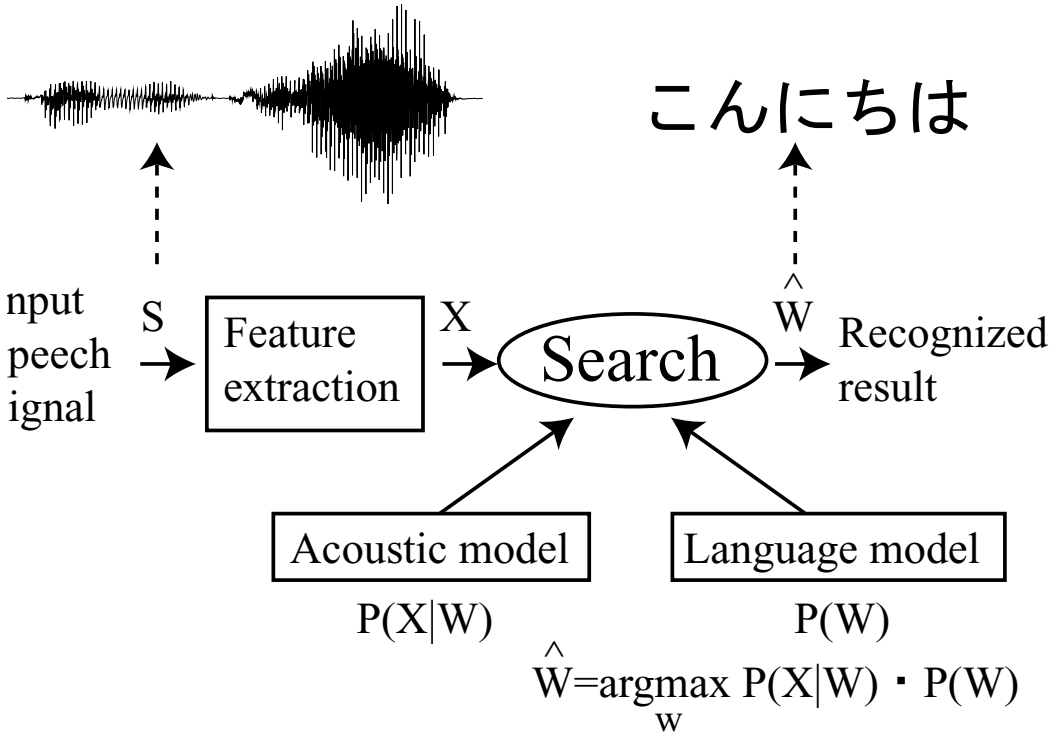

図 2.2: 連続音声認識のフレームワーク

2.2.4　言語モデル

音声認識には音響的な特徴だけでなく, 言語的な知識が用いられる.

例えば, 人間が音声を聞き取る際には, 不明瞭な発音があっても, 文脈上の知識を使って自分で推測して補完している. また, 聞く側が知らない単語が発話された場合は, 明瞭な発音でないと音素を上手く聞き取ることができず, 逆に不明瞭な未知語があった場合は, 既知の単語と聞き間違えてしまう場合がある. つまり, 単に音響的な情報のみで認識しているのではなく, 高度な言語処理を行うことで音声を認識しているといえる. 従って, 機械による認識処理においても音響的処理だけでなく高度な言語処理を行うことで, より自然な推定ができると考えられる. これを実現するのが言語モデルである.

言語モデルで用いられる $P(W)$ は, 与えられた単語列 $w_1 w_2 \ldots w_n$ に対して, その出現確率 $P(w_1 w_2 \ldots w_n)$ を与えるモデルであると考えることができる. 言語モデルとしては様々なものが考えられているが, 特に大語彙連続音声認識においては, 有限状態のネットワーク等で文法を記述することが難しいため, 一般には, コーパスから自動的にモデルを作成する統計的言語モデルのアプローチが広く用いられている.

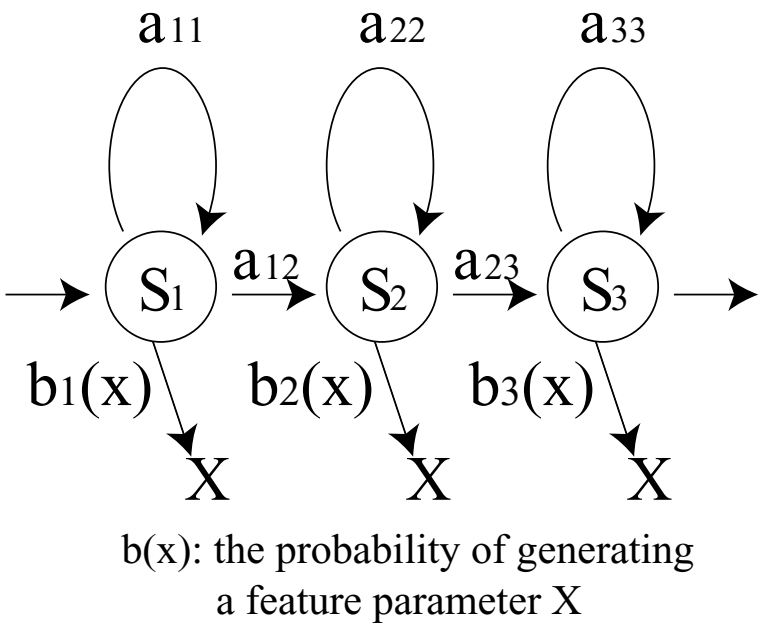

図 2.3: HMM における状態遷移の様子

2.3　N-gram 言語モデル

言語モデルでは, 統計的な推定から次に現れる単語をいくつかに絞って予測することで探索空間を狭めることが重要である. 実際には, 式 2.4 における $P(W)$ を与えるモデルであり, 現在最もよく用いられているのは直前の単語列から次の単語を予測する単語 N-gram と呼ばれるモデルである.

単語 N-gram では, ある単語 w_i が現れる確率をその直前の $N-1$ 単語から予測する. つまり式 2.5 のように, ある単語の出現確率は直前の $N-1$ 個の単語列によってのみ決定されると仮定する.

$$P(w_i|w_1 w_2 \cdots w_{i-1}) \approx P(w_i|\overbrace{w_{i-N+1} \cdots w_{i-1}}^{N-1}) \tag{2.5}$$

この言語モデルのデータ量は, N について指数的に増大するため学習が困難である. また単純に N の値を大きくしていっても, N-gram モデルの精度は向上しないため [5], 現実的には $N=2$ のバイグラム, $N=3$ のトライグラムが主に用いられている. 語彙数を V とすると言語モデルの規模は V^N 程度となる. この規模の確率を数万の語彙に対して計算するには大量のテキストデータが必要になる. 例えばトライグラムの構築には新聞記事 10 年分程度の大規模なデータが必要とされる.

一般的には, このような統計的手法は単語のみを対象とするわけではなく, 一定語彙のシンボルが登場するような現象であればモデル化可能である. 一般的なシンボルを対象として議論する際には, 素性という単語を用いる. 例えば自然言語処理の分野では, 文字 N-gram や品詞 N-gram なども用いられることがある.

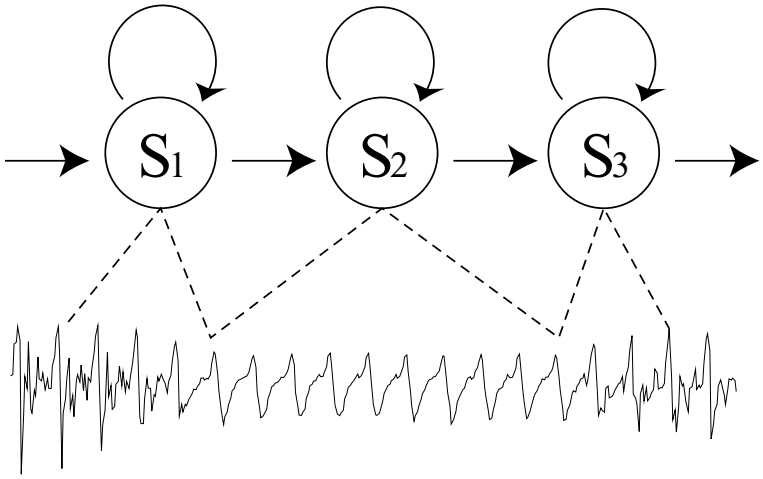

図 2.4: 音素 HMM と対応する音声波形

2.3.1　N-gram 言語モデルの学習

それぞれの N-gram 同時確率は単語 w_i $(i = \{1, 2, \cdots, V\})$ の出現頻度を F_i, また単語列 $\{w_i^n\}_{n=1}^N$ の出現頻度を $C(\{w_i^n\}_{n=1}^N)$ とすると,

$$P(w_i|\{w_i^n\}_{n=1}^{N-1}) = \frac{C(\{w_i^n\}_{n=1}^N)}{C(\{w_i^n\}_{n=1}^{N-1})} \tag{2.6}$$

は, 単語列 $\{w_i^n\}_{n=1}^{N-1}$ が現れたときの条件付確率を表す. $N = 1$ の場合, 式 (2.6) 右辺の分母には, 学習テキスト長 F を与える. この条件付確率のコレクションはそのまま, 学習テキストに対して N-gram 言語モデルの最大の尤度を与える.

このような, 単純な出現頻度に基づくモデル化では, 学習コーパスに出現しなかった単語について確率値ゼロが割り当てられてしまう事になるが, これに対してはスムージングなどの対処が為されることになる. 詳細は 2.3.3 で述べる.

2.3.2　言語モデルの評価

言語モデルの評価は本来, 言語モデルを認識システムに組み込んだときに単語認識率がどの程度改善されたかという尺度で測られるべきである. しかしそのようなテストは大掛かりで大変であるし, 認識性能の差が本当に言語モデルの良さによるものなのかを検証するのも難しい. そのため, 言語モデル単体の評価基準としてパープレキシティと呼ばれる尺度が一般的に用いられている.

言語モデルのパープレキシティとは, ある評価文に対してその言語モデルが次に現れる単語を予測したときに, 予測する単語の平均的な数を表す数値である. 従って, パープレキシティが小さいほど次に現れる可能性のある単語候補を言語的に絞り込んでいることを意味し, 言語モデルの予測精度が高いことを示している.

パープレキシティを求めるためには, まずその言語モデルを評価するための評価用テキストデータを用意する. この評価用テキストデータ $(w_1 w_2 \cdots w_n)$ に対して, 評価したい言語モデルを用いて次式を計算することでパープレキシティが求まる.

$$PP = (P(w_1 w_2 \cdots w_n))^{-\frac{1}{n}} \tag{2.7}$$

この式は, 評価用テキストが現れる確率を単語 1 個当たりの出現確率に変換し, その逆数を用いているので, 結局次に平均何種類の単語候補が予測されているかを計算していることになる. 一般的にパープレキシティが下がると単語正解率は上がる傾向がある.

2.3.3 N-gram 言語モデルの問題点と対策

単語列の頻度に基づき単語の出現確率の推定する N-gram モデルは, パラメータの次元を上げる事無く効果的に推定を行うため, 良く用いられている. その一方, 推定単語の近傍のみの情報を取り扱うと言うモデルの特性上, 問題も持ち合わせている. 以下では, N-gram モデルに関する問題を述べるとともに, 対処する方法に関して説明する.

スパースネスの問題

N-gram の確率値を単純に相対頻度により推定すると, 学習データ中に出現しなかった単語列に対して確率値をゼロにしてしまう. また大規模なコーパスを用いて作成した N-gram 言語モデルにおいても, その大半はほとんどが低いカウント数によるもので, 信頼性の面で問題がある. これらの問題はゼロ頻度問題やスパースネスの問題と呼ばれている [6].

これを補正して N-gram 確率を修正し, 出現確率が 0 でない単語列から, 出現確率が 0 の単語列に対して確率値を分配することを考える. これは一般にスムージング（smoothing）と呼ばれている.

代表的な方法の一つとして, バックオフによるものが挙げられる. これは, 出現確率が 0 でない単語列の出現確率を減少させ（discount）, 変形された単語出現頻度（pseudocount）を実際に用いる.

これは, 学習データ中に存在しない N-gram 確率値を, (N-1)-gram 確率値から推定する手法である. つまり, 学習データにおいて出現頻度が正である単語列はディスカウントされた値をそのまま用い, 0 である単語列はディスカウントによって生じた確率値を低次の言語モデルによる確率値に応じて分配する. ディスカウント係数 $lambda$ を求める方法としては, 加算法, ヘルドアウト推定法, 削除推定法, グッド・チューリング推定法などがある.

バックオフにおいては通常は N-gram に対して (N-1)-gram が用いられるが, [7] では, 予測語に対しそれを予測する文脈の素性それぞれが親変数として与えられ, その選択順番は任意である. すなわち予測文脈長 N とすれば $N!$ のバックオフ経路が与えられる. これをバックオフの一般化と呼んでおり, より多くのバックオフモデルが得られる. このモデル (2-gram) を用いた実験において, 通常のモデルに対してパープレキシティが大幅に低下した.

異なるドメインやスタイルに対する適応

統計的言語モデルは, 学習が行われたドメインや書式などのスタイルなどに依存しており, 認識時の変化に対して弱い [8]. 例えば, 電話での会話を認識する時, そのドメインにおいて 200 万の単語で学習したものの方がラジオやテレビの書き起こしから 1 億 4000 万の単語を使って学習した場合より良い結果となる.

この問題に対処するために, 大規模なコーパスから作成したベース言語モデルと, 小量の分量からなる様々な分野から学習した言語モデルを使って, トピックに応じた適応を行う手法がある [9].

モデルの局所性

N-gram 言語モデルは, ある単語の出現確率を求めるのに非常に近くの単語からの影響しかモデル化できないが, 主語と述語などのように, 現れる位置は離れていても関係の深い組み合わせなどは当然考えられる. これらモデルの局所性に対応したモデルも考案されており, [10] では接続詞間の依存関係を通常の N-gram と同様に局所的制約としてモデル化し, 近接していない語の大局的関係を捉えるため機能語と実質語の影響をモデル化している.

2.4　話し言葉音声認識における取り組み

従来は, 音声認識研究のフィールドとしては書き言葉音声がその対象とされていたが, 近年では話し言葉音声認識への取り組みが増えている. 一般的に話し言葉認識では書き言葉音声認識と比べ様々な点が問題となる [11]. 明示的でない句読点の発見, 様々な不規則発話への対処, 対話での同時発話の認識, 感情やユーザの状況の判断, などである. 特にに本研究では, 講演音声を対象とするので, 様々な不規則発話や明示的でない句読点, すなわちどこで発話が切れるのかという境界発見が問題となってくる.

学習コーパス面においても, 従来は話し言葉スタイルの認識は新聞などの書き言葉コーパスから学習されていたが, 近年になって大規模な日本語話し言葉コーパスが構築された.

本節では日本語話し言葉コーパスである CSJ について説明した後, 話し言葉認識改善への取り組みを, 言語モデルの改良に焦点を当てて紹介する.

2.4.1　日本語話し言葉コーパス (CSJ)

日本語話し言葉コーパス (the Corpus of Spontaneous Japanese)[12] とは, 現代日本語の自発音声 (主に学会講演など) を大量に集めて多くの研究用情報を付加した話し言葉研究用のデータベースであり, 語数にして約 750 万語, 時間にして約 660 時間の音声が含まれている. CSJ は,

- 多数の話者による多少とも自発的な音声を対象としていること
- 豊富な研究用付加情報を提供していること
- 発話スタイルないし自発性に対する評価を与えていること

表 2.1: CSJ が使用する音素セット

a i u e o a: i: u: e: o:
N w y j my ky by gy ny hy ry py
p t k ts ch b d g z m n s sh h f r
q sp silB silE

- XML 文書化されたデータも公開していること

など, 従来の音声データベースにはない多くの特徴を有している.

このコーパスは独立行政法人国立国語研究所と独立行政法人通信総合研究所が推進してきている文科省科学技術振興調整費開放的融合研究制度研究課題「話し言葉の言語的・パラ言語的構造の解析に基づく『話し言葉工学』の構築」プロジェクトの一環として構築されている. このプロジェクトの目標は自然な話し言葉を工学的に処理するための基盤技術を開拓することにおかれているが, CSJ はそのために必要不可欠なデータベースとして位置づけられており, その構築作業は主として国立国語研究所が分担している. CSJ は 2004 年 6 月に公開された.

また, CSJ の中には, CSJ で学習された音声認識用の音響モデルと言語モデルが含まれている. 音響モデルは時間にして 486 時間分の 2496 講演から, 言語モデルは語彙にして 6.6M を含む 2592 講演から学習されている. 音素体系は表 2.1 に示す 42 種類である. ここで, q は促音に伴う無音, sp は音声中の短い無音である. silB は発話の先頭の無音, silE は発話の終端の無音であり, 発話は基本的に 500ms 以上の無音区間で区切ったものと定義している. N は撥音, a:〜o:は長母音を表す. 形態素は国立国語研究所で定義された短単位 [13] に基づいている.

2.4.2　話し言葉認識に対する言語モデル改良の取り組み

フィラーの話者性を考慮した言語モデル

[14] では, 学習時に, 学習コーパス中の全話者 (391 人) に対し, フィラーの種類を用いて話者クラスタリングを行い, クラスタごとのフィラーモデルを作成している. そして, 認識時には, 最初にいったん話者の音声を認識してその話者のクラスタを同定し, そのクラスタのフィラーモデルを用いて改めて認識を行う, 2 パス処理をしている.

話者クラスタリングについては, 以下の手順で行っている.

- 類似したフィラーをひとつのカテゴリにまとめる (人手)
- フィラーの分布を仮定し, 話者ごとの距離を定義する (Kullback-Leibler Divergence を利用)
- 話者のクラスタリングを行う (クラスタ数 10)
- クラスタごとの言語モデルを作成

これらの処理をした後, 6 名の評価話者に対する認識実験を行ったが, 成果を上げることができなかったと報告されている.

不規則発話の言語モデルへの組み込み

[15] では, 事前に用意された話題に関しての電話の会話音声に関するコーパスである Switchboard(SWBD)[16] に関して, 不規則発話要素 (DF) に注目して統計的言語モデルを構築した. これは, DF の直後の単語は, DF の無い単語列から予測する事でさらに正確なものになると同時に, DF 自体は通常の単語と同様にコンテクストに応じた確率で出現する, と言う仮定を置くものである.

言語モデルをモデリングする際に, 以下のような規則でおこなう.

- DF の後に登場する単語の N-gram は,DF をコンテクストに含めない
- 簡単のため, DF はフィラー (filled pauses), 言い直し (repetitions), 脱落 (deletions) のみとする
- DF は文脈上の条件として確率的に割り当てられた事象として, 普通の単語予測と同じように N-gram で予測される

結果は, パープレキシティにおいて, ほとんど改善が見られなかったが, ある特定の場所においてはパープレキシティ改善が見られた. 具体的には (境界ではなく) 節の中間にあるフィラーに対しては, 本手法を用いる事で直後の語予測に効果があった. 著者らは, 言語的境界に出現するフィラーは, 次の単語の目印になっており, これが最初の仮定を違反しているとしている. 境界や節の中間といった場所に応じて言語モデルを適応させていく手法が効果があると考えられる.

言語モデルの話者変動に対する教師なし適応

[17] では, 言語モデルの教師なし適応を用いて改善を図っている. これは, 話し言葉において文末表現等で発話の傾向が話者毎に違う事を考慮したもので, 話者性に対する適応手法として, 認識結果を直接用いて適応する手法, 発話文単位で類似テキストを選択しそれを用いて適応する手法, の二つを上げている.

認識結果を用いる手法では, 一度認識した結果を用いてバックオフ単語 3-gram を作成する事で, 実際の話者特性に対応する. 認識結果より得られたモデルと通常のモデルを線形補間することで作成される. 両者の補間係数は別に用意された development-set (評価テキストの一部:評価には用いない) を用いて推測される.

類似テキストを選択する方法では, 候補である各学習データに対し前出の方法 (一度認識した結果を用いる方法) により作成したバックオフ言語モデルを用いてパープレキシティを計算し, ある閾値より低いものを「類似テキスト」とし, 改めて学習を行う. 得られた言語モデルと通常の言語モデルを線形補間して各話者に適応した言語モデルを作成する.

二つの手法を統合した結果, パープレキシティにおいて 68.18%から 53.20%へ改善し, 単語誤り率において 33.8%から 28.7%へ改善した.

2.5 まとめ

本章では, 大語彙連続音声認識システムの概要について説明した. さらにシステムにおける言語モデルの役割と, 頻度に基づく統計的言語モデル故の問題点について言及した. さらに話し言葉の認識における問題点を取り上げて, 言語モデルに注目した認識改善の取り組みをいくつか取り上げた.

次章では, 境界情報を用いた言語モデルに関する基礎的な考えを述べ, さらに先行研究について紹介する.

第3章

境界情報を用いた言語モデル

3.1 はじめに

我々は人間が音声の認識に利用すると考えられる発話境界情報に注目して, 言語モデルを高度化する試みを行ってきた. これは, 境界を跨ぐ場合と跨がない場合の単語遷移の様子の違いに着目したもので, 両者の単語 N-gram を個別に取り扱うものである.

まず, ATR501 文連続音声コーパスのアクセント句境界を利用するもの [1] では, 品詞遷移に着目することでコーパスサイズの不足に対処することを行った. 次に, 文節境界に着目することで, 大規模新聞コーパスの利用を可能とした [18]. これらは朗読調音声を対象としたものであったが, 日本語話し言葉コーパス（Corpus of Spontaneous Japanese, CSJ）を利用し, 話し言葉音声を対象とした研究を進めてきた. 既に, アクセント境界を利用することで, 言語モデルの性能を向上し得ることを示している [4]

本章では, まず境界の有無により言語モデルを作り分けることの妥当性を確かめるために, 境界付近での種々の性質に関して述べる. 続いて, 境界情報を利用した言語モデルに関しその構成法を説明し, さらに上記の取り組みに付いて紹介する. 最後に問題点に付いて整理する.

3.2 境界とは

境界の定義は様々であるが, 一般的には, ある何らかの特徴を持つセグメント間自体の事を指すと認識する事ができる. ここでは, 代表的なものとしてアクセント句境界と文節境界, また文節境界の一種である統語境界について述べる.

3.2.1 アクセント句境界

日本語の単語は音の高さが「高→低」へと変化する箇所[1]が１つあるか, または１つもないかのいずれかである. この「高→低」と高さがシフトする部分をアクセント核と呼ぶ. 文中では助詞なども含めて, この高さがシフトする部分をひとつだけ含むような区分の方法がある.

物理的な意味としては, アクセント句は基本周波数の山にほぼ相当する. アクセント句境界は, その境界にあたる韻律イベントである. 図 3.1 は, 音声の基本周波数パターンとアクセント句, アクセント句境界を示した図である. もちろん, アクセント句境界の位置は同じ文を発声しても発話の仕方によって異なる. 例えば図 3.1 の例では, 「あらゆる現実を」までを一つのアクセント句として発声することも可能である.

アクセント句境界の検出法としては, ヒューリスティックに基づくもの [19], HMM を用いて各フレームが句境界である尤度を算出する方法など [20], またモーラを単位としたモーラ遷移確率モデルによってアクセント型を HMM でモデル化して, アクセント句境界を検出するもの [21] などがある.

[1] これをアクセント核という.

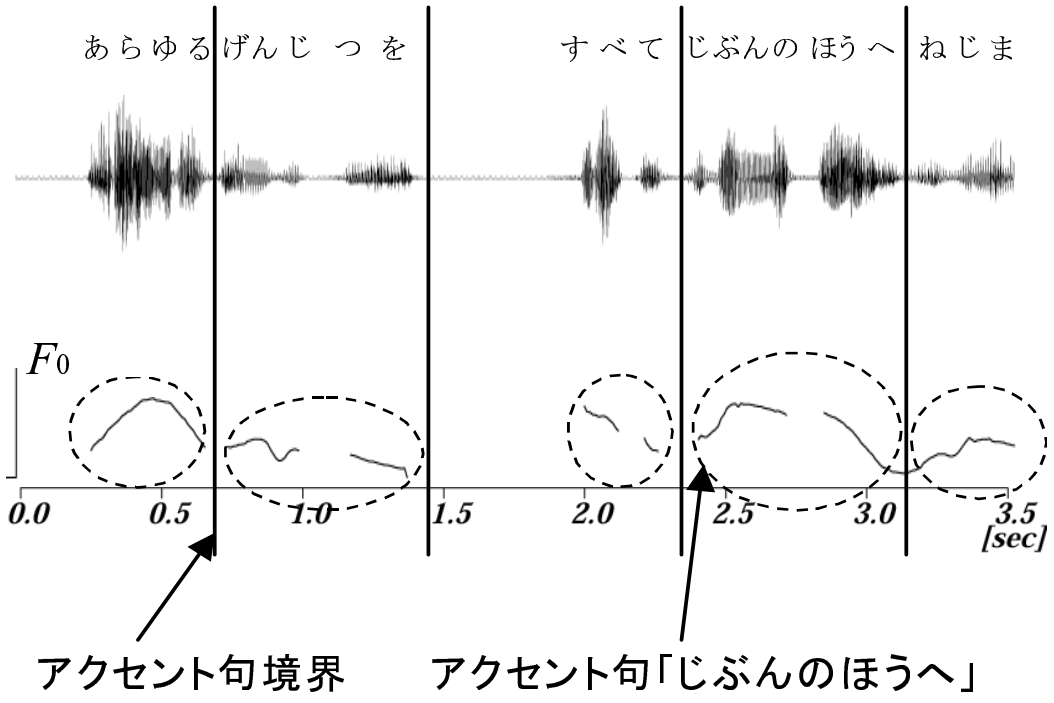

図 3.1: アクセント句境界

3.2.2 文節境界

文節とは, 橋本進吉によって定義された日本語における文法的な単位であり, 一つの概念語 (content word), あるいは概念語＋機能語 (function word) で構成される.

日本語話し言葉コーパスである CSJ においては, 係り受け構造情報の基本単位として文節を用いており, 文節に対する独自の認定基準が使用されている. 基本的には, 以下に示す本則 [A], [B], [C] の順に適応され, さらにいくつかの例外規則が存在する.

- 本則 [A] 助詞・助動詞連続
  助詞・助動詞連続の後で切る.

- 本則 [B] 助詞・助動詞を伴わない自立語
  助詞・助動詞を伴わない自立語については, 主語・主題の後や連用／連体修飾成分の後, 接続詞や感動詞の後, などで切る.

- 本則 [C] 体言連続
  以上の規則に該当しない場合, 一部が連体修飾成分を受けている体言連続の後, 同格・言い替えの体言連続, 並列された語などの関係を切る.

表 3.1: 統語境界の分類

タイプ	形態
A 類	～ながら,～つつ,～たり,～なく
B 類	～と,～ば,～たら,～なら,～て,～てから, ～ても,～ず,～ず (に)
C 類	～から（理由）,～ので,～のに,～けど,～けれど, ～が,～し,～で,～まして
D 類	終助詞,～です,～ます,～でした,～, ました,～ん

3.2.3　統語境界

従属節の終わりに現れるものである. 日本語の従属節は, 主節に対する従属度の観点から, 大きく A, B, C の 3 種類に分類されている [22]. A から C にいくほど主節との関連性が低くなり, 独立度が高くなる. 分類を表 3.1 に示す.

また,[23] では, 明確な文末形式で終わるものを D とし, 統語的境界とフィラーの出現確率の関係について述べている. 統語的境界は $A \sim D$ まであり, $A \rightarrow D$ となるに従って「深く」なる. これらの境界に現れるフィラー頻度について調べた結果, 統語的に「深い」境界にフィラーが多く現れるとしている. ただし,D の文境界に関しては,C より出現率が低かった. その原因として, 文境界でのポーズにより後続発話のプランニングができるのでフィラーの必要性が下がる事などを挙げている.

3.3　境界の有無による言語的性質の違い

3.3.1　句境界情報を用いた連続音声認識

[24] では, 連続音声認識システムの探索効率と認識精度を向上させるために, アクセント句境界情報を連続音声認識の性能改善に利用する方法を提案し, その中で, アクセント句境界周辺における認識尤度の変化について言及している.

図 3.2 において, 図中の点は正解となる認識経路のスコア, 実線では固定ビーム幅によって枝刈りされた際にビーム内に残った候補の最低スコアである. 図 3.2 によれば, 正解経路のスコアは単語が遷移する際に一時的に落ち込んでいることが分かる. これは単語が遷移するときに言語モデルの尤度が加えられるためであるが, 多くの単語遷移が考えられるとき言語モデルによる予測が困難になり, 言語モデルにおける尤度が低くなる. そのため, ここで誤った認識が発生しやすい. そこで, あらかじめ与えられていたアクセント句境界情報を元に, 句境界がそこにある場合はビーム幅を拡げ, 句境界が存在しない句内ではビーム幅を徐々に狭めていくという動的な制御を行っている. アクセント句境界がある場合は, 単語遷移の結びつきが弱くなると考えられるためである.

また, アクセント句境界が存在する場合は前後の音素環境を考慮した音素モデルを用いなかった. これは, アクセント句境界では前後の音素の影響は受けにくいと考えられ, 逆に正確な認識を阻害すると考えられるからである.

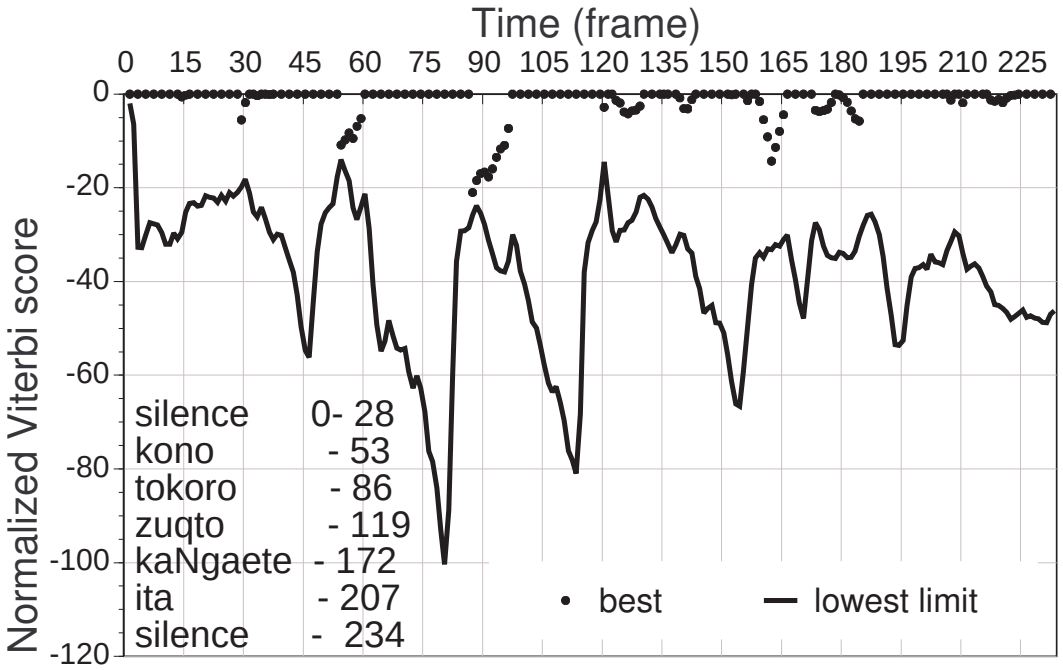

図 3.2: 正解パス（点）と最低スコアを持った候補（実線）の Viterbi スコア

ビーム幅の動的な制御によって同じ認識率に対して計算時間と消費メモリの大幅な削減が達成され, また音素モデルの選択によって単語認識率, 文正解率が改善された.

3.3.2　アクセント句境界有無による品詞遷移傾向の違い

[1] では, アクセント句境界の有無に応じた言語的性質の違いとして, 品詞遷移傾向を調査している. 具体的にはアクセント句内部および句境界における品詞の遷移の頻度をカウントし, 確率をそれぞれ算出した. 調査対象は, 朗読調音声コーパスである ATR503 文 [25] である. 品詞体系としてはテキストを chasen で形態素解析した結果を用いた. 調査結果の中から「名詞」「動詞」「助詞」「副詞」について 4 品詞間の句内部および句境界での遷移確率を示したものを表 3.2, 3.3 に示す.

3.4　境界情報を用いた言語モデル構築の基本的アイデア

境界情報を用いた言語モデルを構成する際の基本的アイデアとしては, 境界を跨ぐ時と跨がない時の単語の遷移傾向の違いを織り込んで言語モデルを作成し, 認識時に境界予測尤度と組み合わせる事で精度の高い認識を実現するものである. これを図 3.3 に示す. 図では境界情報を用いなかった時の 2-gram 確率 $P(Y|X) = \frac{1}{2}$ であるが, 最終遷移で境界を跨いだ場合／跨がない場合に分けてカウントを行い, 集合の母数と現れる単語の傾向から境界を跨がない場合の 2-gram 確率が高くなっている.

表 3.2: ATR 句内部遷移 [1]

		遷移先			
		名詞	動詞	助詞	副詞
遷移元	名詞	8.9	5.2	67.5	0.1
	動詞	6.2	12.7	43.8	0.0
	助詞	6.8	47.9	36.9	0.3
	副詞	2.5	15.0	60.0	0.0

表 3.3: ATR 句境界遷移 [1]

		遷移先			
		名詞	動詞	助詞	副詞
遷移元	名詞	71.1	13.4	1.4	2.8
	動詞	85.6	5.7	1.1	4.0
	助詞	51.1	34.3	0.2	6.1
	副詞	59.6	28.8	0.0	1.4

これら境界を跨いだ語によって作成された N-gram と境界を跨がなかった語によって作成された N-gram を, 境界予測尤度 Pb を用いて線形補間する. なお境界を跨がない確率は $(1 - Pb)$ として表される. 図 3.3 の場合では $Pb = \frac{1}{3}$ (つまり, 境界ではなさそうな確率が高い) であり, 最終的な結果は $P(Y|X) = \frac{19}{38} > \frac{1}{2}$ (通常の N-gram による確率) となり, 境界情報を活用する事で予測確率が大きくなっている.

ここで境界尤度 Pb は, 言語的な情報や韻律情報などを用いた様々な境界予測手法により予測される事になる.

3.5　アクセント句境界検出を用いた言語モデルの高精度化

[1] では, アクセント句境界情報を利用して bigram 言語モデルの高精度化が実現された. これは, 入力音声の韻律分析によってアクセント句境界を抽出し, 図 3.4 に示すようにアクセント句境界の有無に応じた 2 個の言語モデルを用意することにより音声認識タスクにおける言語モデル部分の高精度化を行ったものである.

この手法の狙いは, アクセント句境界の有無によって日本語の言語的な性質が異なり, それが単語遷移確率の値となって現れることを期待し, これを利用することである.

学習・評価実験には ATR の音韻バランス文が使用され, アクセント句境界の抽出手法としては [21] と同じものが用いられた. その際のアクセント句境界の抽出精度は検出率 57%, 挿入誤り率 24% であったが, ATR の音韻バランス文の学習・評価テキストに大して, およそ 5% 程度のパープレキシティ低下（改善）が報告されている.

また, ATR の音韻バランス文での品詞遷移におけるアクセント句境界の有無のカウント

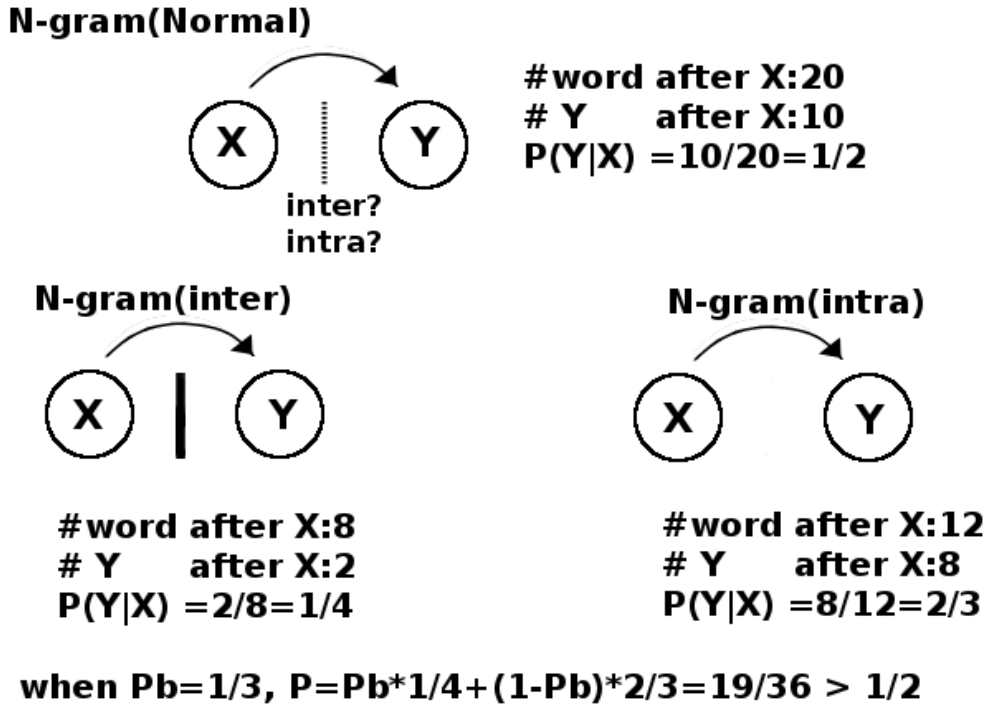

図 3.3: 境界情報を用いた言語モデル構築の基本アイデア

の比率を求め, 6 年分の新聞記事コーパスを用いたベース言語モデルに対して単語カウントを分配するという手法を用いた場合では, およそ 10% 程度のパープレキシティ低下が報告されている.

3.6　文節境界を用いた言語モデルの高度化

[3, 18] では, 境界として文節境界を用いる事で, 言語モデルの改善を図っている. これは, アクセント境界が実際の音声発声によるもので, その情報をテキストから得ることは難しく, そのため学習テキストが不足してしまう, という事態に対処したものである. 文節境界は文法的にも発音的にも一つの単位として認識されうるため, テキストのみから境界が検出できる.

境界尤度は単語系列より求められるが, ここで N-gram モデルの場合, N 以上の単語系列からより良い境界予測を行う事がポイントとなる. 予測単語数による境界の予測能力を調べた予備実験によると, 8 割以上の確率で境界があるかどうか予測できる場合を正解とした時, 予測単語数 2 の時約 84%だったのに対し, 予測単語数 4 では約 93%となり, より長い単語系列により境界予測の精度が良くなる事がわかった.

文節境界をまたぐ言語遷移と, 文節境界をまたがない言語遷移のそれぞれから構築され

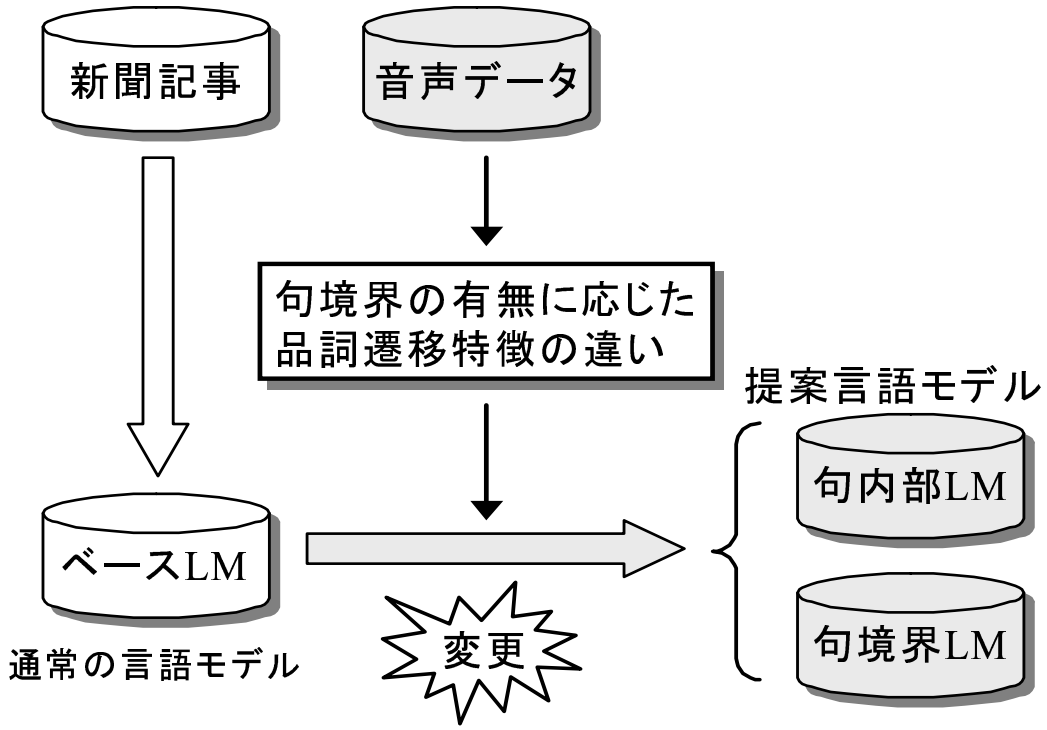

図 3.4: アクセント句境界の有無に応じた言語モデルの使い分け

た二つの言語モデルを以下の式により統合する.

$$P_{total}(w_n|w_{n-N+1}^{n-1}) = pb * P_{inter}(w_n|w_{n-N+1}^{n-1}) + (1-pb) * P_{intra}(w_n|w_{n-N+1}^{n-1})) \qquad (3.1)$$

ここで,pb は学習テキストにより算出された, 境界の出やすさ (尤度) を表す.

[18] では, 毎日新聞 1 年分から学習されたモデルにより, パープレキシティにおいて改善がみられた. また [3] においては音声認識において評価された. 毎日新聞 1 年分による結果ではベースラインとの間に平均 8%程度の改善がみられたが, 毎日新聞 3 年分による学習では, あまり明確な違いはみられなかった. これは学習コーパスが大きくなる事により, ベースラインの言語モデルに暗黙の形で境界情報が含まれてしまったのではないかと考察している.

3.7 話し言葉におけるアクセント境界を用いた言語モデルの高度化

上記の研究は主に朗読調の発話に適用されたものである.[4] では, さらに, 日本語話し言葉コーパス (CSJ) に, 韻律的特徴から判断されるアクセント境界を境界として適用している. アクセント句境界における韻律境界らしさの確率 $P(b)$ を求める方法として, 以下の 2 つが組み合わせて用いられている.

1. 境界付近の韻律的特徴 ([4] では,F_0 のみ)
2. 先行する境界からの距離

1. に関しては, 連続するモーラ区間中の F_0 系列を特徴量ベクトル $\mathbf{x}$ とし, クラス b が $\mathbf{x}$ を出力する確率を混合正規分布を用いて表している. 2. は, 時間・モーラ数・単語数に依存する混合ポワソン過程により表現される.

評価実験においては, CSJ のコアデータを用いる事により境界尤度の学習が行われた. また, 言語モデルの学習においては, 境界を跨ぐ／跨がない 2 つの 3-gram 言語モデルを, コアデータだけで学習するには十分な量が得られないため, まずは境界を跨ぐ遷移と境界を跨がない遷移に分けて, 品詞 3-gram を作成した. これを, CSJ を用いて学習した (通常の) 単語 3-gram 言語モデルに適用し, 境界を跨ぐ／跨がない 2 種類の単語 3-gram を作成した. ディスカウント法は Witten-Bell 法を用いた.

評価は, 句境界推定を行う提案手法, CSJ に付与された句境界情報の正解を与えた提案手法, さらに CSJ 全体で学習された (通常の)3-gram 言語モデルの間でパープレキシティを比較した. 評価データセットはコアデータ中の学会講演・再朗読が用いられた.

結果は, ベースラインに比べて提案手法を用いることにより, 若干の改善が確認された. さらに, 句境界の正解を与えた場合の方が性能が良かった.

[4] では, フィラーにより句境界が検出されない例を挙げ, そのために, 句境界の推定を行った場合が正解を与えた場合に比べ性能が悪くなったとしている.

3.8 まとめ

本章では, 境界情報に関する基本的考察と, その特性について述べた. また, 文節境界, アクセント句境界といった境界に対して境界検出と境界情報をどのように利用して言語モデルを構築するか説明した.

先行研究の結果から, 学習時と認識時で境界の条件を同一にした方が結果が良い事, 境界予測尤度が高い方が必ずしも最終的な結果に結び付かないことがわかった. また, 話し言葉における不規則発話要素の影響で境界検出精度が下がった結果, 最終的な認識結果が悪くなる事があった.

次章では, 本章の結果をふまえて境界検出手法や言語モデルの特性などに注目して評価実験を行っていく.

第4章

言語モデルの高度化に関する検討

4.1 はじめに

本章では, まず書き言葉と同様に話し言葉コーパスにおいても品詞の遷移傾向に違いが見られるか, 境界を跨ぐ場合と跨がない場合に分け調査を行う.

続いて, 品詞による境界予測を用いた言語モデルを作成し, 単語による境界予測の場合と比較を行う. これは, 品詞による境界予測が特に小さいコーパスにおいて有効に行われるとの考えによる. 境界予測を行う場合と, 境界の正解を与える場合で言語モデルの性能がどう変化するか確認する.

次に, 言語モデルのバックオフを考慮し, 境界を跨ぐ／跨がないモデルのそれぞれにおいてどの長さの N-gram が多く用いられているか, またパープレキシティはどのようになっているか確認する事でモデルの改善に繋げる.

最後に, 通常のモデルと境界モデルの統合モデルを提案し, 検証実験を行う.

4.2 話し言葉における品詞遷移傾向の調査

[1] では, 書き言葉コーパスである ATR503 文において, 境界を跨ぐ／跨がない場合に分けて品詞の遷移傾向を調査した結果, 両者に差がある事を結論づけ, これが境界を用いた言語モデルの利用モチベーションの一つとなっている. そこで, 本節では話し言葉における場合について調査を行った.

4.2.1 話し言葉における品詞

日本語話し言葉コーパス CSJ における品詞遷移の状況について, 文節境界をまたぐ時とまたがない時にわけて, 傾向を調査した. 特に, それぞれの品詞ごとでどのような遷移をしているか, 境界を越える時と境界内での遷移に分けて調査した. CSJ における品詞のリストを表 4.1 に示す.

表 4.1: CSJ に登場した品詞/活用形リスト

名詞	代名詞	形状詞
連体詞	副詞	接続詞
感動詞	動詞	形容詞
助動詞	助詞	接頭辞
接尾辞	記号	言いよどみ

4.2.2 実験

文節境界をまたぐ遷移, またがない遷移, の 2 つに分けて, 形態素の品詞遷移頻度の傾向をカウントした. 形態素としては, CSJ に定められている短単位を用い, また文節境界も CSJ

表 4.2: CSJ 文節内部での品詞遷移

		遷移先			
		名詞	動詞	助詞	副詞
遷移元	名詞	19.5	7.5	55.4	0.0
	動詞	3.6	0.8	42.9	0.0
	助詞	1.4	48.3	35.6	0.0
	副詞	0.7	70.3	19.8	0.7

表 4.3: CSJ 文節境界での品詞遷移

		遷移先			
		名詞	動詞	助詞	副詞
遷移元	名詞	36.0	9.4	13.5	3.3
	動詞	71.0	1.65	2.5	1.5
	助詞	39.2	22.9	0.6	4.9
	副詞	38.1	16.7	0.3	6.4

の定義に従った [13]. 調査した文書は, CSJ 中の「学会講演」987 ファイルである.

4.2.3 結果

調査結果から, その一部を示す. 表 4.2 に文節内部遷移の結果を, 4.3 に文節境界での遷移結果を示す.

結果は遷移元の品詞ごとに正規化されており, 全遷移先の確率の和を取ると 100%となる.

4.2.4 考察

話し言葉コーパスである CSJ においても, 文節境界を跨ぐ時と跨がない時で, 品詞遷移傾向に大きな違いが見られることがわかる. これは, N-gram 言語モデルの最終遷移において, 境界遷移の有無により言語モデルを 2 つ作る事の妥当性を示している.

また, 本結果より, 品詞を用いた境界予測には効果があると考えられる. 次章では品詞遷移による境界予測を用いた結果と単語遷移による境界予測を用いた結果を比較する.

4.3 境界予測単語／品詞数とコーパスサイズによる違い

日本語話し言葉コーパス CSJ[12] を用いて言語モデルの構築と評価を行う. 境界情報としては, アクセント句境界を用いる事も考えられるが, 今回はより多くのデータが利用可能であるという観点から, 文節境界を用いて言語モデル構築を行う.

言語モデル構築においては, まず予測語の直前に境界を跨ぐ語遷移と跨がない語遷移に分けて, それぞれから言語モデルを学習する. さらに, 先行する H 個の単語／品詞から境界尤度を算出し[1], 式 3.1 によって 2 つの言語モデルから得られる次単語の確率を統合する.

4.3.1 実験

CSJ のコアに含まれる学会講演を用いたもの (モデル 1) と, 学会講演・模擬講演を用いたもの (モデル 2) を用いて 3-gram 言語モデルを構築した. それぞれのサイズを表 4.4 に示す.

表 4.4: 言語モデルサイズ

モデル 1	語数：260k, 語彙数：8k
モデル 2	語数：440k, 語彙数：16k

境界予測モデルは, それぞれの言語モデル構築時に用いたコーパスを利用した. 直前の 3-4 単語／品詞から境界の有無に応じて頻度を計数し, 境界尤度を算出した. なお, 予測単語／品詞数が多い場合, 評価テキスト中にそのような単語／品詞列が現れない場合がある. その時は単純に低次の境界予測モデルによって予測を行う.

評価は, 提案言語モデルと, 従来の言語モデル間で, パープレキシティの比較を行った. さらに, testset 中に境界 (正解) 情報を与えることで, 境界を跨ぐ言語モデルと境界を跨がない言語モデルの一方を選択的に用いる場合の評価を行った. 評価テキストは CSJ の学会講演の文章（open, close）を用いた.

評価テキストの詳細について表 4.5 に示す. testsetC は close data, testsetO は open data である.

表 4.5: 評価テキスト

testsetC	男性, 語数 2616, 語彙数 442
testsetO1	男性, 語数 2930, 語彙数 599
testsetO2	女性, 語数 2478, 語彙数 593

加えて, 予測する境界尤度のパープレキシティを, モデル 2 において算出した. これは, 境界尤度予測のあいまいさを図る指標で, 1 から 2 までの数値をとる. 1 に近いほど, 境界を跨ぐ／跨がない, がより明確に予測されているといえる.

4.3.2 結果

実験によって得られた結果を表 4.6, 4.7 に示す. H は境界予測に用いた品詞数を示す. pp はパープレキシティ, $diff\_pp$ はベースラインとの差を示す. また improve は改善率で,

[1] H 個の単語／品詞の直後に, 境界が出現する／しない頻度を用いる.

$100 * \frac{diff\_pp}{baseline}(\%)$ である．また境界尤度予測のパープレキシティについては表 4.8 に示す．

4.3.3 考察

モデル 1, モデル 2 それぞれの場合において, 単語境界予測の場合が品詞境界予測の場合に比べ改善している．しかし open data においては, 単語境界予測の場合の方が境界予測パープレキシティは若干高く, 境界予測精度としてはあいまいである．正解情報があり完全に境界が予測できている場合にもパープレキシティが悪化しており, 境界尤度予測の精度以外の面で考察が必要である．

一方, この場合の close data では若干だが改善がみられている．この理由として, 境界の正解を利用して片方の言語モデルを利用する場合, close data では 3-gram でのヒット率が高く, バックオフによる 1-gram のヒット率が低いため, 高次 N-gram の持つ高い確率値の影響が出ているのではないかと考えられる．

従って, 次章では境界 (正解) 情報を与えたときに, どの N-gram がヒットするか傾向を調べ, それぞれの N-gram に関するパープレキシティを比較することで, 以上の考察を検討する．

4.4 N-gram ヒット率とパープレキシティ

境界を用いた言語モデルの場合でも, テキスト評価中にヒットする N-gram がない場合, バックオフにより (N-1)-gram を利用して確率を算出することになる．本章では, 境界の正解情報を与えた評価テキストを用いて, 境界を跨ぐ／跨がない言語モデルの片方を選択的に利用し, その場合の N-gram 利用率とそれぞれのパープレキシティを調査する．さらに通常の言語モデルに対しても, 境界の有無別に使われた N-gram の頻度を N 別に調査し, 比較する．

4.4.1 実験

前節で用いたモデル 2 を使って実験を行う．評価項目は境界を跨ぐ／跨がない言語モデルのそれぞれにおける N-gram(N=1, 2, 3) 利用率とパープレキシティである．通常の言語モデルに対しても, 使われた状況別 (境界を跨ぐ／跨がないとき) に N-gram 利用率を計測する．パープレキシティも同様に測定する．評価テキストは表 4.5 と同様である．

4.4.2 結果

境界利用モデルにおける調査結果を表 4.9 に示す．また, 通常の言語モデルにおける結果を 4.10 に示す．各 N-gram ヒット数の割合は全ヒット数における割合である．またパープレキシティは, それぞれの N-gram 確率値とヒット数に基づいて算出されている．

表 4.6: モデル 1 の性能評価

testsetC

		品詞予測	単語予測
baseline			10.74
H=3	pp	17.52	25.04
	diff_pp	-6.77	-14.30
	improve	-63.02	-133.06
H=4	pp	17.83	25.50
	diff_pp	-7.09	-14.76
	improve	-65.96	-137.38
正解あり	pp		9.40
	diff_pp		1.34
	improve		12.50

testsetO1

		品詞予測	単語予測
baseline			145.55
H=3	pp	140.15	131.40
	diff_pp	5.40	14.16
	improve	3.71	9.73
H=4	pp	140.98	132.45
	diff_pp	4.58	13.10
	improve	3.14	9.00
正解あり	pp		164.16
	diff_pp		-18.61
	improve		-12.78

testsetO2

		品詞予測	単語予測
baseline			141.22
H=3	pp	144.43	134.84
	diff_pp	-3.22	6.38
	improve	-2.28	4.52
H=4	pp	145.55	135.28
	diff_pp	-4.34	5.94
	improve	-3.07	4.20
正解あり	pp		179.69
	diff_pp		-38.48
	improve		-27.25

表 4.7: モデル 2 の性能評価

testsetC

		品詞予測	単語予測
baseline			12.38
H=3	pp	20.55	31.27
	diff_pp	-8.17	-18.89
	improve	-65.98	-152.58
H=4	pp	20.79	31.96
	diff_pp	-8.40	-19.58
	improve	-67.88	-158.11
正解あり	pp		10.76
	diff_pp		1.62
	improve		13.06

testsetO1

		品詞予測	単語予測
baseline			163.81
H=3	pp	157.45	150.65
	diff_pp	6.37	13.17
	improve	3.89	8.04
H=4	pp	157.78	151.86
	diff_pp	6.04	11.96
	improve	3.69	7.30
正解あり	pp		181.67
	diff_pp		-17.85
	improve		-10.90

testsetO2

		品詞予測	単語予測
baseline		173.70	173.70
H=3	pp	176.94	171.65
	diff_pp	-3.24	2.05
	improve	-1.87	1.18
H=4	pp	177.67	172.14
	diff_pp	-3.97	1.56
	improve	-2.29	0.90
正解あり	pp		221.20
	diff_pp		-47.50
	improve		-27.35

表 4.8: 境界予測パープレキシティ

		品詞予測	単語予測
testsetC	H=3	1.21	1.10
	H=4	1.22	1.10
testsetO1	H=3	1.24	1.31
	H=4	1.24	1.30
testsetO2	H=3	1.22	1.27
	H=4	1.22	1.26

4.4.3 考察

close data においてはほぼ N=3 で N-gram がヒットしている．また境界を跨がない言語モデルのパープレキシティが小さく, そちらが多くヒットしている状況がパープレキシティの改善に繋がっていると考えられる．

open data においては testsetO1 と testsetO2 で似たような傾向を示している．境界を跨がないモデルでは, N が減少するに従ってヒット率は減少し, パープレキシティは増加している．これに対し, 境界を跨ぐモデルにおいては N=3 と N=1 の割合が多く, また N の違いによるパープレキシティの値にかなり大きな開きがある．通常の言語モデルでは, 表 4.10 より, 境界を跨ぐ場合, 跨がない場合にかかわらず使用される N の数が少なくなるほど利用率が少なくなる傾向がみられる．

これより, 境界情報を利用したモデルにおいては, パープレキシティの高い N=1 のモデルを利用する状況が多いことが, 全体としての性能の悪化を招く要因の一つであると考えられる．

4.5 境界モデルと通常モデルを用いた融合モデルの提案

前節までの結果から, 境界を跨ぐ／跨がないモデルにおいて, 使用される N-gram の N(3,2,1) 別の結果においてはパープレキシティはベースラインより低いものの, 高次の N-gram においてテストセットと適合するものが少ないため, 低次の N-gram がベースラインのモデルに比べて使用されている率が高いことが全体として性能の悪化につながっていると考えられる．

これは, 境界情報を利用したモデルが境界を跨ぐ／跨がないに応じて 2 つの言語モデルに振り分けられる事で作成されるため, 全てのコンテクストから作成した通常のモデルに対し, スパースネスの問題を持っているためと考えられる．

本節では, 上記の問題を解決するため, そもそも境界を利用したモデルにおいてどのような状況下が最もその特性を活かす事になるのか考察し, それをもとに通常の言語モデルとの補間モデルを提案し, 実証を行う．

表 4.9: ヒット率とパープレキシティ評価 (境界情報を利用したモデル)

(a) testsetC

割合 (%)			
	N=3	N=2	N=1
inter	41.93	0.04	0.00
intra	58.03	0.00	0.00
perplexity			
	N=3	N=2	N=1
inter	48.00	10.94	*
intra	3.65	*	*

(b) testsetO1

割合 (%)			
	N=3	N=2	N=1
inter	19.49	4.85	17.82
intra	30.61	15.63	11.60
perplexity			
	N=3	N=2	N=1
inter	257.73	326.65	47960.30
intra	4.27	23.64	4893.43

(c) testsetO2

割合 (%)			
	N=3	N=2	N=1
inter	16.95	4.16	17.88
intra	28.17	19.73	13.12
perplexity			
	N=3	N=2	N=1
inter	223.69	1154.54	83488.60
intra	4.45	33.44	3366.70

表 4.10: ヒット率とパープレキシティ評価 (通常の N-gram)

(a) testsetC

割合 (%)			
	N=3	N=2	N=1
inter	41.93	0.04	0.00
intra	58.03	0.00	0.00
perplexity			
	N=3	N=2	N=1
inter	59.79	9.81	*
intra	3.97	*	*

(b) testsetO1

割合 (%)			
	N=3	N=2	N=1
inter	19.49	10.96	11.71
intra	30.60	17.71	9.56
perplexity			
	N=3	N=2	N=1
inter	300.13	1687.10	49187
intra	5.21	19.01	10153

(c) testsetO2

割合 (%)			
	N=3	N=2	N=1
inter	17.07	12.07	9.85
intra	28.17	22.44	10.41
perplexity			
	N=3	N=2	N=1
inter	254.65	3067.71	101670
intra	5.14	27.10	6020.44

4.5.1 境界情報を使ったモデルの活かされる条件

境界情報を利用した言語モデルにおいて, あるコンテクスト w(3-gram なら, 前の 2 単語) から次単語 x を予測する確率について考察する.

まず通常の N-gram 言語モデルについて考える. コンテクスト w の次に現れる単語の総数 (種類×それぞれの数) を N, 単語 x の総数を N_x とする. このとき, 予測尤度 P_0 は次のように表される.

$$P_0(x|w) = \frac{N_x}{N} \tag{4.1}$$

次に, 境界を跨ぐモデルについて考察する. コンテクスト w の次に現れる境界を跨ぐ単語の総数 (種類×それぞれの数) を $N_{(inter)}$, 単語 x の内, 境界を跨ぐものの数を $N_{x(inter)}$ とすると,

$$P_{inter}(x|w) = \frac{N_{x(inter)}}{N_{(inter)}} \tag{4.2}$$

である.

ここで, コンテクスト w の次に現れる単語の総数 (種類×それぞれの数) における境界を跨ぐ単語の 割合 を P'_b, コンテクストの次に現れる単語 x の内, 境界を跨ぐものの 割合 を P_{bx} とすると, 次のように表される.

$$P_{inter}(x|w) = \frac{P_{bx}N_x}{P'_bN} = \frac{P_{bx}}{P'_b}P_0(x|w) \tag{4.3}$$

また境界を跨がないモデルは次のようになる.

$$P_{intra}(x|w) = \frac{(1-P_{bx})N_x}{(1-P'_b)N} = \frac{1-P_{bx}}{1-P'_b}P_0(x|w) \tag{4.4}$$

よって, 境界情報を利用した言語モデルの予測する確率は, 以下のように表される.

$$P(x|w) = p_b\frac{P_{bx}}{P'_b}P_0(x|w) + (1-p_b)\frac{1-P_{bx}}{1-P'_b}P_0(x|w) \tag{4.5}$$

ただし, p_b は別の方法より求められる境界予測尤度である.

式 4.5 より, $P_{bx} = P'_b$ の時,

$$P(x|w) = p_bP_{base}(x|w) + (1-p_b)P_{base}(x|w) = P_{base}(x|w) \tag{4.6}$$

となり, 境界モデルとベースラインが等しくなる.

以上の議論を整理する.

- 境界の存在確率の偏りと予測単語が境界の次に存在する確率の偏りの傾向が似ている時には, 境界情報を利用する意義は少ない

よって, このような条件下においては, スパースネスに強い通常の言語モデルを使う事で対処する事を考える.

表 4.11: 提案モデルにおける境界モデルの使用率 (パーセント)

	testsetC		testsetO1		testsetO2	
	正解あり	境界予測	正解あり	境界予測	正解あり	境界予測
T=0.01	81	81	64	52	62	51
T=0.1	48	48	35	35	34	28

4.5.2 実験

実際に P_{bx} や P'_b を求めるのは困難であるので, 代替手段を用いる. 式 4.3, 4.4 により, $P_{bx} = P'_b$ のとき $P_{inter} = P_{intra}$ であることがわかる. この条件において通常の言語モデルと境界を用いたモデルが等しくなる. 本実験に置いては, より通常の言語モデルの利用割合をもたせ, 式 4.7 の条件において, 通常の言語モデルを用いる.

$$|P_{inter} - P_{intra}| < threshold \tag{4.7}$$

学習には, 前節で用いたモデル 2 を用いる. テストセットにおいて境界情報を与えたもの (正解あり) と境界予測を行うものの評価を行った. 境界予測には品詞／単語を用いた. なお評価テキストは表 4.5 と同様である. $threshold$(閾値) は 0.01 と 0.1 の時の評価を行った.

4.5.3 結果

提案モデルにおける境界モデルの占める割合を表 4.11, 結果を表 4.12 に示す. ベースラインとしては通常の N-gram を用いた. 従来手法は通常の N-gram モデルとの補間を行わない境界情報利用モデルである. T は, 境界を跨ぐ／跨がないモデルの差の閾値を示し, $|P_{inter} - P_{intra}|$ がこの値より小さければ, 通常のモデルを使う事を示している.

4.5.4 考察

通常のモデルとの補間を行うことにより, 性能の改善がみられている.

オープンデータに関しては, 特に境界の正解情報を与えた場合がベースライン, 従来手法のそれぞれに比べ良い結果を示してしており, 境界情報を適切に利用し, スパースネスの問題にも対処できていると考えられる. クローズデータに関しては, 従来手法の性能が良いが, これは, クローズデータにおいては境界利用モデルにおいて性能が良い高次のモデルがもともと利用されているためだと考えられる.

一方, 境界予測を行った時には性能に差が見られている. testsetO1 では $T = 0.01$ の時が良く, testsetO2 では $T = 0.1$ の時が良い. さらに testsetO2 では, T に関して境界予測を行う場合と境界の正解を与えた場合で傾向が逆になっている. 境界予測尤度が正解情報に近いか, という面での境界予測精度が影響しているのではないかと考えられる.

表 4.12: 提案言語モデルの性能評価 (T:閾値, pos:品詞予測, word:単語予測)

(a) testsetC

		正解あり	pos(H=3)	pos(H=4)	word(H=3)	word(H=4)
baseline		12.38				
T=0.01	pp	11.14	19.12	19.27	27.3	27.79
	diff_pp	1.24	-6.74	-6.89	-14.92	-15.41
	improve	10.06	-54.46	-55.65	-120.49	-124.45
T=0.1	pp	11.68	15.49	15.56	18.91	19.1
	diff_pp	0.7	-3.11	-3.18	-6.53	-6.72
	improve	5.62	-25.11	-25.65	-52.77	-54.31
従来手法	pp	10.76	20.55	20.79	31.27	31.96
	diff_pp	1.62	-8.17	-8.4	-18.89	-19.58
	improve	13.06	-65.98	-67.88	-152.58	-158.11

(b)testsetO1

		正解あり	pos(H=3)	pos(H=4)	word(H=3)	word(H=4)
baseline		163.81				
T=0.01	pp	147.61	157.24	157.78	152.69	153.61
	diff_pp	16.2	6.57	6.03	11.13	10.2
	improve	9.89	4.01	3.68	6.79	6.22
T=0.1	pp	156.06	158.74	159.02	154.84	155.14
	diff_pp	7.75	5.07	4.79	8.97	8.67
	improve	4.73	3.1	2.92	5.47	5.29
従来手法	pp	181.67	157.45	157.78	150.65	151.86
	diff_pp	-17.85	6.37	6.04	13.17	11.96
	improve	-10.9	3.89	3.69	8.04	7.3

(c) testsetO2

		正解あり	pos(H=3)	pos(H=4)	word(H=3)	word(H=4)
baseline		173.7				
T=0.01	pp	162.43	172.15	172.82	172.12	172.7
	diff_pp	11.27	1.54	0.88	1.58	1
	improve	6.48	0.89	0.51	0.91	0.58
T=0.1	pp	168.71	171.4	171.96	171.52	171.82
	diff_pp	4.99	2.3	1.74	2.18	1.87
	improve	2.87	1.32	1	1.26	1.08
従来手法	pp	221.2	176.94	177.67	171.65	172.14
	diff_pp	-47.5	-3.24	-3.97	2.05	1.56
	improve	-27.35	-1.87	-2.29	1.18	0.9

4.6 まとめ

本章では, 話し言葉コーパスにおける品詞の境界を跨ぐ／跨がない時の遷移傾向の違いについて調査し, 傾向に差がある事が分かった. その結果より, 品詞を用いた境界予測を提案し単語による予測と比較を行った. 結果は, 境界予測精度が良い場合でも全体としてパープレキシティが低下してしまう事があったため, 実際に使用される N-gram の N の数に注目し, それぞれの出現割合とパープレキシティを評価した. 境界を利用したモデルではパープレキシティの高い低次のモデルが多く使われていることがわかったため, 境界情報を活かしつつスパースネスの問題に対処するため通常の N-gram モデルとの補間モデルを提案した. その結果, 性能の改善がみられ, 特に境界の正解を与えた場合において最も良い結果がみられた. 境界予測を行った場合においては, 境界予測精度などの点からさらに検討が必要である.

第5章

結論

5.1 本研究のまとめ

本研究では, 境界情報を利用した言語モデルの高度化に関して, 境界予測手法, コーパスサイズ, バックオフの影響による N-gram の利用率とパープレキシティの傾向などの面から調査, 考察を行った. また境界情報の有効利用と, スパースネスの問題に対処するため, 通常の言語モデルとの補間を行う新しい言語モデルを提案した.

品詞／単語を用いた境界予測による実験の結果から, 従来手法においては, 境界を正確に予測する事がパープレキシティの低下といった言語モデルの性能改善に必ずしも繋がっていなかった. このため, 境界の有無別に, バックオフの影響を考慮した N(1,2,3) 別の利用率とパープレキシティ調査を行う事によって, 境界情報利用モデルの利用状況について考察した. 通常の言語モデルに比べ, 境界情報を用いた言語モデルでは, 精度の高い高次の N-gram はあまり用いられず, 精度の低い低次の N-gram が多く用いられている事が分かった. これは, 境界を跨ぐ／跨がないで分けた事によるスパースネスの問題が現れた結果だと考えられる.

そのため, 高次のモデルでのヒット率が高いベースモデルと, 使用されるコンテクストをより反映した境界情報利用モデルの融合を提案した. まず理論的に考察を行う事で, 実際にどのような状況で境界情報を利用したモデルが効果を示すのかを明らかにした. 効果が発揮される条件下においては境界モデルを利用し, そうでない場合はスパースネスに強いと考えられる通常のモデルを利用することで, より効果的な境界情報の利用を図った. その結果, ベースライン, また従来手法に比べて, 言語モデルの性能面での向上がみられた.

5.2 本研究の問題点と今後の課題

まず, 通常モデルと境界モデルの融合モデルにおける閾値に関して, 今回は事前に割り当てたものであったが, どのように最適な閾値を割り当てるか, また認識結果を使って動的に変えていくのか, といった検討が必要である. さらに境界予測を行った際には, 境界予測手法によって性能にばらつきがみられている. 境界予測尤度が正解にどれだけ近いのか, といった精度の点などから検討していく必要がある.

本研究においては, 3-gram において境界モデルが作成されている. 境界の有無はその直前の単語に最も影響されると考えられるので, 2-gram による検証も必要である. また 3-gram においても予測すべき単語の直前の境界の有無のみが考慮されているため, もう一つ前の単語の直前の境界の有無がどのような影響を与えているのか調査する必要がある. また今回はコーパスサイズが他の書き言葉コーパス等に比べて小さい状況での実験だった. その状況では提案モデルにより性能が向上することを示したが, コーパスサイズを増加させていったときにどのようになるかを検証する必要がある. 提案モデルにおいては, 音声認識実験による評価も行う必要がある.

5.3　今後の展望

5.3.1　話し言葉特有の要素に注目した言語モデリング

話し言葉特有の要素であるフィラー, 言い直し, 言い淀み等, それぞれの特性に注目した言語モデリングを行う事で, 性能の改善が行える可能性がある. 例えばフィラーの扱いに関しては, 文の途中に登場し予測確率を下げる存在であると同時にある種の予測器としての役割を持っている. 境界の深さ等に応じてフィラーを透過する, 透過しない言語モデルを組合せて利用する事により, フィラーの特性を活かしつつより良い予測を行う事が期待される.

5.3.2　様々な境界を利用した言語モデル

従来, アクセント句境界や文節境界が境界として用いられてきたが, 明示的ではなくても別の境界情報を用いて言語モデルの高度化を図る事が考えられる. 例えば, CSJ の長単位の間を境界として利用し, さらに文節境界と併用する事で, 境界の種類に応じた言語モデル選択の可能性が広がり, より細かい特徴把握により, 性能改善が可能になるのではないかと考えられる.

謝辞

本研究は様々な人からアドバイスを頂き完成しました. 特に, 広瀬啓吉教授の御指導には感謝しております. 丁寧でありながら適切なアドバイス, また時には厳しいご指摘を受ける事で, 研究の方向性に道標を付ける事ができました. また, 峯松信明助教授には, 研究に関する基本的姿勢を教えて頂きました. さらに研究をどう見せるか, どのように伝えるか, と言う面で多くを学びました.

研究室のメンバーには, 研究におけるアドバイスのみならず, 遊びなどの面でもお世話になり, 充実した生活を送る事ができました. この研究室で切磋琢磨できたことで, 2 年間の生活が実りある深いものとなったと感じております.

みなさまに深く感謝の意を示させて頂きます.

2009 年 2 月 4 日

細田聖人

参考文献

[1] 寺尾真, 峯松信明, 広瀬啓吉. アクセント句境界情報を利用した N-gram 言語モデルの高精度化. 電子情報通信学会技術研究報告, SP2001-101, 2001.

[2] Keikichi Hirose and Makoto Terao Nobuaki Minematsu. Statistical language modeling with prosodic boundaries and its use for continuous speech recognition. In *Proc. ICSLP*, Vol. 2, pp. 937–940, 2002.

[3] Sungyup Chung, Keikichi Hirose, and Nobuaki Minematsu. N-gram Language Modeling of Japanese Using Bunsetsu Boundaries. In *Proc. Interspeech*, pp. 993–996, 2004.

[4] 上西康太, 広瀬啓吉, 峯松信明. アクセント句境界を考慮した言語モデルの日本語話し言葉コーパスへの適用. 日本音響学会秋季講演論文集, pp. 67–68, 2007.

[5] M. Weintraub, Y. Aksu, S. Dharanipragada, S. Khudanpur, H. Ney, J. Prange, A. Stolcke, F. Jelinek, and L. Shriberg. *Fast Training and Portability* 1995 language modeling summer research workshop. Technical report, Johns Hopkins University, 1996.

[6] 北研二. 確率的言語モデル, 第 3 章. 1999.

[7] Jeff A. Bilmes and Katrin Kirchhoff. Factored language models and generalized parallel backoff. In *HLT/NACCL*, 2003.

[8] Ronald Rosenfeld. Two decades of statistical language modeling: Where do we go from here? In *Proceedings of the IEEE*, Vol. 88, pp. 1270–1278, 2000.

[9] Reinhard Kneser and Volker Steinbiss. On the dynamic adaptation of stochastic language models. In *Proc.ICASSP93*, Vol. 2, pp. 586–589, 1993.

[10] R.Isotani and S.Matsunaga. A stochastic language model for speech recognition integrating local and global constraints. In *Proc. IEEE-ICASSP*, 1994.

[11] Elizabeth Shriberg. Spontaneous speech: How people really talk and why engineers should care. In *Proc. INTERSPEECH*, pp. 1781–1784, 2005.

[12] 前川喜久雄. 「日本語話し言葉コーパス」付属ドキュメント, 「日本語話し言葉コーパス」の概観. 2004.

[13] 小椋秀樹. 日本語話し言葉コーパスの構築法, 第 3 章 形態論情報. 2006.

[14] 板垣貴裕, 篠田浩一, 嵯峨山茂樹. 話し言葉音声の認識における間投詞の話者性を考慮した言語モデル. 第 2 会 話し言葉の科学と工学ワークショップ講演予稿集, pp. 79–84, 2002.

[15] A. Stolcke and E. Shriberg. "Statistical Language Modeling for Speech Disfluency". In *ICASSP*, pp. 405–408, 1996.

[16] J.J. Godfrey, I.C. Holliman, and J. McDaniel. Swichboard;telephone speech corpus for research and development. In *Proc. ICASSP*, pp. 517–520, 1992.

[17] 南條浩輝, 河原達也, 山田篤, 内元清貴. 講演音声認識のための言語モデルの教師なし適応. 情報処理学会研究報告. SLP, 音声情報処理, 2002.

[18] 鄭聖曄, 広瀬啓吉, 峯松信明. 文節境界情報を利用した N-gram 言語モデルの高精度化. 情報処理学会研究報告. SLP, 音声言語情報処理, pp. 13–18, 2003.

[19] Keikichi Hirose, Atsuhiro Sakurai, and Hiroyuki Konno. "Use of Prosodic Features in the Recognition of Continuous Speech". In *ICSLP1994*, Vol. 1, pp. 149–152, 1994.

[20] 花沢俊之, 阿部芳春, 中島邦男. ピッチパタンの統計モデルに基づく句境界情報を利用した文節スポッティング. 日本音響学会誌, Vol. 55, No. 1, pp. 23–31, 1999.

[21] Keikichi Hirose and Koji Iwano. Detection of prosodic word boundaries by statistical modeling of mora transitions of fundamental frequency contours and its use for continuous speech recognition. In *Proc. IEEE ICASSP*, pp. 1763–1766, 2000.

[22] 南不二男. 現代日本語の構造. 大修館書店, 1974.

[23] 渡辺美知子, 伝康晴, 広瀬啓吉, 峯松信明. フィラーの出現確率予測における節の種類と後続節長. 日本音響学会秋期講演論文集, pp. 319–320, 2005.

[24] Shi wook Lee, Keikichi Hirose, and Nobuaki Minematsu. "Efficient search strategy in large vocabulary continuous speech recognition using prosodic boundary information". In *ICSLP2000*, Vol. 4, pp. 274–277, 2000.

[25] 阿部匡伸, 匂坂芳典, 梅田哲夫, 桑原尚夫. 研究用日本語音声データベース利用解説書 (連続音声データ編). 1990.

発表文献

[1] 細田聖人, 広瀬啓吉, 峯松信明, "話し言葉認識における文節境界情報を用いた言語モデルに関する検討", 日本音響学会秋季講演論文集, pp.95–96, 2008.

[2] 細田聖人, 広瀬啓吉, 峯松信明, "境界情報を用いた言語モデルの高度化に関する検討", 日本音響学会春季講演論文集, 2009, 発表予定.