

木構造を用いた音韻連鎖統計モデル

7 B-5

田本 真詞

伊藤 克亘

田中 穂積

(東京工業大学)

1はじめに

連続音声認識において音韻認識誤りを補正するための言語情報のひとつとして、音韻連鎖統計モデルが広く用いられている。音韻連鎖の統計モデルとは、情報発生源である文章や会話の音韻連鎖の内在的構造を、実際に収録された文章や会話などの有限のトレーニングデータから推定するものである。このような統計モデルとしてN-gramによる音韻連鎖の確率モデルが音声認識に有効に働くことが知られている[1]。しかしN-gramモデルでは、音韻連鎖長Nの値が増加するにつれ、モデルの分解能が向上する反面、モデルを推定するために必要な標本数が減少し、個々のモデルの推定精度が低下するためモデル全体の統計的信頼性が損なわれる問題点があった。

我々は、統計的な信頼性を損なうことなく分解能、推定精度を向上させながら音韻連鎖の生成確率をモデル化するために、コンテキストに応じて参考する連鎖の長さを動的に変化させる、木構造を用いた音韻連鎖統計モデル(Vari-gram)を提案した。Vari-gramモデルは、木構造を用いて音韻連鎖のコンテキストを表現し、情報量の大きなコンテキストを選択的に分割、成長させることでモデルの長さを増加させ、分解能の高いモデルを実現する。この時、モデルの信頼性の低下を生じさせるような標本数の少ないコンテキストの成長は抑制され、モデルの信頼性が維持できる。

また、従来から用いられているN-gramに対し、タスクとの適合性を示すcoverageや、一文あたりの生成確率、認識タスクの複雑度を示すパープレキシティーの点で優れていることを実験によって検証した[2]。

2 Vari-gram モデル

本稿では、Vari-gramの持つ特質のうち、トレーニングデータからのタスク抽出能力やデータの減少に対するrobustness(頑健性)を検証する。

トレーニングデータの減少が統計モデルに及ぼす影響としてタスクの局在化があげられる。統計モデルによって得られるのは、対象コンテキストのトレーニングデータにおける出現頻度から求められる推定値 \hat{P} であり、コンテキストの真の生成確率 P に対して常に何らかの誤差を含んでいる。この推定値は、音韻連鎖長の増加やトレーニングデータの減少などによってコンテキストあたりの標本数が減少すると真の生成確率 P に対する誤差が増大し、モデルの統計的な信頼性が低下する。このようなモデルは、トレーニングデータのタスクに過適応なためトレーニングデータに存在しないタスクに対応することができず、モデルとしての信頼性は低い。

はじめに述べた通りVari-gramは、情報量の大きなコ

ンテキストをトレーニングデータから抽出することで、分解能、推定精度の高い、信頼性のあるモデルを生成するが、これは、モデルの頑健性やタスク抽出能力にも反映されると考えられる。

3 実験

トレーニングデータのサイズによってモデル化の変化する様子をcoverage、モデルの平均音韻連鎖長、一文あたりの平均生成確率、テストセットパープレキシティーの各項目についてVari-gramとN-gram($N = 2, 3, 4, 5$)の間で比較した。実験に用いたテキストデータベースは、1982年の日本経済新聞の記事37日分の15,207文、文節数は145,718、音韻数で1,385,082である。また、生成されたモデルの評価用として、同新聞の1日文の記事、571文、6260文節、54733音韻を推定用のデータとは別に用意した。

4 結果

トレーニングデータの量を増減させた場合の、各評価の変化を図1~4に示す。各グラフにプロットされた値は、図中に示す通りトレーニングデータに含まれる音韻数を増減した時のVari-gram(折線)、N-gram(点)の変化を表す。N-gramは、グラフの左側から $N = 2, 3, 4, 5$ の結果を表している。

coverage

統計モデルによって、テストセットの生成確率を算出する場合、参照されたモデルのうち、生成確率が0とならなかったモデルが全体に占める割合を%で表す。この値が大きいほどモデルが訓練データのタスクを良く表現できている。

グラフ(図1)では、トレーニングデータのサイズの減少によるVari-gramのcoverageの悪化は、N-gramに比較して小さい。これは、Vari-gramが標本数の多いコンテキストをモデル化するため、データサイズ減少の影響を受けにくいものと考えられる。

音韻連鎖の平均長

音韻連鎖の平均長は、各々のモデルの音韻連鎖長の平均値である。グラフ(図1)では、トレーニングデータが減少するにつれ、同じモデル数のN-gramとVari-gramでは、N-gramの方が値が大きい。ただし、N-gramのモデルには推定精度の低いコンテキストが含まれていると考えられ、次に示す音韻の平均生成確率に影響を及ぼすものと思われる。

音韻の平均生成確率

音韻連鎖が $P = p_1, p_2, \dots, p_n$ である一文に含まれる音韻の平均生成確率を次の式で定義する。

$$P(P) = \prod_{i=1}^n P(p_i | p_{i-N+1}, p_{i-N+2}, \dots, p_{i-1})^{\frac{1}{n}}$$

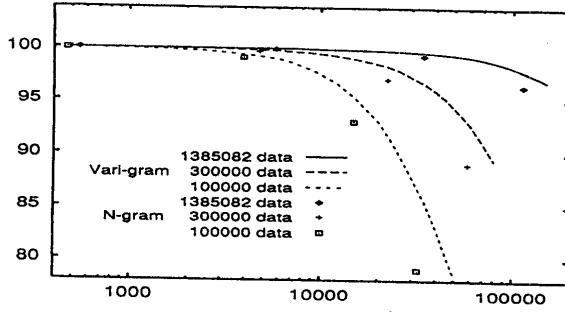


図 1: coverage

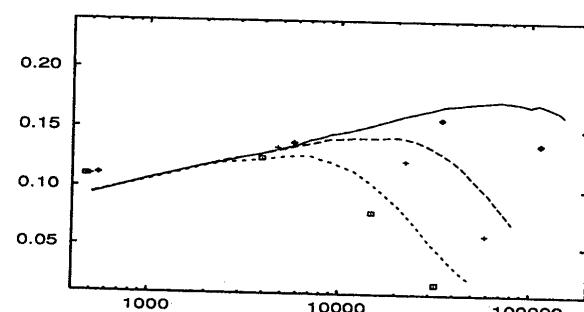


図 3: 音韻連鎖の生成確率

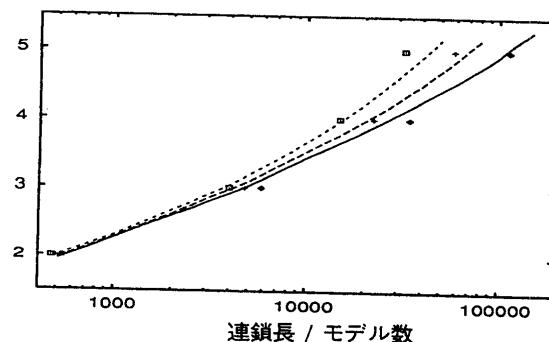


図 2: モデルの平均連鎖長

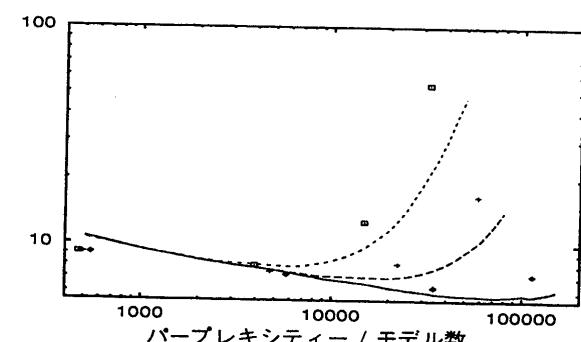


図 4: パープレキシティー

これは、文中の音韻の生成確率の平均値である。したがって、推定精度・信頼性の高い言語統計モデルほど「日本語らしさ」を持った文に対して高い値を示す。

平均生成確率(図 3)は、データサイズの減少によって大きく低下するが、Vari-gram は N-gram より優れた値を示す。これも N-gram が coverage の悪化などでモデルの信頼性が低下するのに対し Vari-gram は、頑健であることを示している。

パープレキシティー

パープレキシティーは、統計モデルを連続音声認識処理で用いる確率的言語情報としたときの、認識のタスクの複雑さを評価する尺度のひとつである。これは、認識対象のエントロピーをもとに算出される、情報理論的な意味での平均分岐数である。パープレキシティーは、値が大きくなるほど複雑さが増大し、認識が困難になる。

データの減少に従い、パープレキシティーは急激に増加する(図 4)。この場合も、coverage や平均生成確率に優る Vari-gram の方が値が低く、タスクの複雑さが軽減されている。

5 結論

実験結果から Vari-gram は N-gram に対して coverage、平均生成確率、パープレキシティーの項目で優れており、特に coverage とパープレキシティーは、トレーニングデータが減少した時に差が広まる傾向にある。このことから Vari-gram は、トレーニングデータにタスクの局在化が生じてもモデルの一般性を失わないタスク抽出的一般性とパープレキシティーを大きく上昇させない頑健性を持つものと結論できる。

ある大きさを持つトレーニングデータに対してどの規模までコンテキストを選択、Vari-gram モデル化するのが妥当なのか、判定が難しい。今回用いた評価尺度の中には、モデル数の変化に対して極値を取るものがあり、このような判断の基準として用いることができるかどうか、今後の課題となる。

謝辞

本稿で使用した日本経済新聞の記事に関するテキストデータベースは、NTT 情報通信研究所メッセージシステム研究部から提供して頂きました。貴重なデータを使用させて頂いたことを深く感謝いたします。

また、日頃から研究について御助言、御討論頂く田中研究室の皆さんに感謝いたします。

本研究は、筆者が(株)エイ・ティ・アール自動翻訳電話研究所で実習生として在籍中に、音声情報処理研究室長の嵯峨山茂樹氏より Vari-gram の名称と本モデルの基本概念について御教示頂いたものを、同様の概念を持っていた[3] 東京工業大学工学部情報工学科田中研究室の伊藤克亘と共に発展充実させ、名称を同じく Vari-gram とし今日に至ったものです。

参考文献

- [1] 川端、花沢、伊藤、鹿野、「HMM 音韻認識における音節連鎖統計情報の利用」、信学技法、SP-89-110 (1990)
- [2] 田本 真詞、伊藤 克亘、田中 穂積、「木構造を用いた音韻連鎖統計モデル」、電子情報通信学会技術報告 SP92-93 (1992)
- [3] 伊藤克亘、「日本語の統計的な振舞いを利用した連続音声認識」、修士論文、東京工業大学、(1990)