確率的シソーラスと文書クラスタに基づいた トリガー言語モデルの拡張による音声認識

トロンコーソ・アラルコーン・カルロス (110089)

北陸先端科学技術大学院大学 情報科学研究科

2003 年 2 月 14 日

キーワード： 言語モデル, 音声認識, 確率的シソーラス, トリガー言語モデル, EM アルゴリズム.

1 はじめに

近年、音声認識において最も幅広く使用されている言語モデルは、n-gram モデルである。n-gram は、文中で互いに近接している単語間の依存性 (近距離依存性) をモデル化するのに効果的な手法である。しかし、n-gram は過去の $n-1$ 単語に制限された単語の履歴に依存しているため、出現位置の離れた単語間の依存性 (長距離依存性) のモデル化に適していない。

n-gram の制限を回避するために提案された手法の一つとして、トリガー言語モデルがある。トリガー言語モデルは、直前に出現した少数の単語を記憶しておくキャッシュモデルと同様に、キャッシュコンポーネントを利用している。さらに、トリガーペアと呼ばれる意味的に関係がある単語ペアの集合も利用している。トリガーペアは、膨大な記事コーパスから平均相互情報量を用いて構築される。 キャッシュ内の全単語、全モデルだけでなく、トリガーペアを通して関連した全単語は、文中に出現する確率が高いと考えた手法である。

トリガーモデルの欠点は、キャッシュモデルとパフォーマンスがよく似ていることにある。なぜなら、最良なトリガーの大部分は、自己トリガー (self-triggers) であり、語幹に関連したトリガーであるためである。

単語間の相互関係が改善できれば、音声認識において、より重要な効果をもたらすトリガーペアを得るのは可能であると考えられる。

本研究で提案する拡張トリガー言語モデルは、まずトリガーペアに替わり、使用された関連単語のペアの確率的シソーラスを用いる。さらに、文書クラスタから、関連単語を抽出して、キャッシュ内に組込む。

2　提案手法

2.1　概念

本研究での提案手法は、引用する単語がそれ自身と関連を持つように処理された二つの異なる情報源を利用している。これらは、確率的シソーラスと文書クラスタである。

単語で構成される確率的シソーラスと、関連する「後置詞＋単語」の組合わせは、意味的なクラスにまとめられる (例えば、電車, バス,... ↔ に乗る, の運転手,...)。それぞれのクラスは、2 セット” 主要単語” に分類される。単語が互いに意味的な関係を持ち、関連単語のセットになっている。すなわち、” 主要単語” に関連した単語が後置詞を通じてセットになる。これは、統計学的な構文解析ツールと EM アルゴリズムを基としたクラスタリングを用いることにより、大量のテキストコーパスから自動的に生成され、トリガーペアよりも強い関連を持つ単語間の総合的かつ意味的な関係を獲得する。

文書クラスタは、これらの文書に現れるであろう単語に従って、それらの確率分布をもとに同様の内容を持つ文書クラスタで構成される (例えば、文書 573, 文書 947,... ↔ 電車, 駅, 線,...)。それらは、同じテキストコーパス (5 年分の日本語の新聞) から EM クラスタリングの手法によっても作成され、同様の文書のセットで重要なトピックを示す単語を特定できる。

キャッシュには、後置詞を除く、主要単語と関連単語が加えられ、確率的シソーラスにおける最も適切なクラスの単語もキャッシュに加えられる。さらに、それらがキャッシュに含まれていなければ、文書クラスタの最も適切なクラスタからも単語がキャッシュに加えられる。

トリガーモデルと提案手法の主な相違点は、以下に示す通りである。

まず、モデルが異なるデータを使用している。トリガーペアは、同じコンテキストの中で現れる類似性のある関連単語の組合わせである (例えば、教育 → 大学)。一方、確率的シソーラスは、意味的なクラスの中で、後置詞を通じて関連単語の組合わせを分類する。また、単語の用法の相違を反映する (例えば、” ダイエー” デパートの名前であったり、野球チームの名前であったりする)。

さらに、提案モデルは、強い相関を持つ名詞と動詞 (例えば、ビール ↔ 飲む)、名詞と名詞 (例えば、巨人 ↔ 投手) など、単語間のより良い総合的で意味的な関係をモデル化する。

2.2　定式化

本研究の提案手法は、新しい言語モデルによって算出されるスコアを用いた音声認識システムが出力する N-best 仮説をリスコアする。

提案した言語モデルのスコアは、以下の式 (1) で示される拡張キャッシュコンポーネントのスコア ($S_{extended}(W)$) と認識器が出力するベースラインの言語モデルのスコア ($S_{baseline}(W)$)

からなる。

$$S(W) = S_{extended}(W)^{\lambda} S_{baseline}(W)^{1-\lambda} \tag{1}$$

式 (1) において、λ は重み係数、W は処理されている文 (単語列) である。

このように、ベースラインのモデルによってモデル化された近距離依存性をうまく利用し、提案モデルが捕らえる長距離依存性を加えることが可能である。

拡張キャッシュコンポーネントのスコアは、文中に含まれる全単語のキャッシュスコアを正規化したものである。

$$S_{extended}(W) = \prod_{i=1}^{n} (S_{cache}(w_i))^{\frac{m}{n}} \tag{2}$$

式 (2) において、n は文中に含まれる単語数であり、m は N-best の文の平均単語数である。

ある単語のキャッシュスコアは、キャッシュ内のユニグラム確率によって定義される。式 (2) に示すように、キャッシュに単語が存在するとき 0 に近い値、そうでなければ ε となる。

$$S_{cache}(w_i) = \begin{cases} \frac{N_{cache}(w_i)}{Cache\ \ Size} & N_{cache}(w_i) \neq 0 \\ \varepsilon & \text{otherwise} \end{cases} \tag{3}$$

式 (3) において、$N_{cache}(w)$ は、キャッシュに w が出現した数を表す。

3 実験結果

男性話者 2 名による、71 文からなる二つの異なるテストセットでの実験を行った。実験データは、読売新聞の教育に関する記事で構成される。

音声認識システム Julius3.1 は、モデルのリスコアをする N-best の仮説を出力するために使用した。このシステムは、2 パス探索を行い、第一パスで bigram、第二パスで trigram を用いて、再探索、再評価を行う。本実験では、$N = 100$(100-best の仮説出力のリスコアを行う問題) に設定した。

この時、100-best の認識率は 91.35%であり、この実験における理論的な最高認識率とする。

確率的シソーラス中の 2500 クラスから選択される重要なクラス数は 5、また、各クラスから選択される主要単語数、及び関連単語数はそれぞれ 5 である。300 の文書クラスタから選択される重要なクラスタ数は 1 であり、各クラスタから選択されるの重要単語は 5 である。従って、提案モデルによるキャッシュサイズは、標準のキャッシュベースモデルのサイズの 56 倍になる。

キャッシュコンポーネントのみによるモデルと、確率的シソーラスと文書クラスタに基づく拡張トリガーモデルの認識精度 (単語正解精度) を、λ の値を 0 から 1 の間で、0.05 ずつ増加させて計算した。また、基本となるキャッシュサイズは、5,10,25,100,250 と 500 である。

結果として、提案手法 (キャッシュサイズ=25) は N-best の最高認識率を基準としたとき、従来手法と比べて、13.5%の誤り削減率を実現した。

4 結論

本研究では、拡張トリガー言語モデルを提案した。従来のトリガーモデルとは異なり、提案手法は、以下の二つの情報源、すなわち、確率的シソーラスと文書クラスタを基本としている。前者は、文中の単語間の意味の依存関係と同様の文法関係を得ることができ、後者は、現在の会話の話題についての情報を与える。

実験では、二つの情報源から抽出された関連単語が、言語モデルとして良い制約を与え、音声認識性能の向上に有効であることを示した。