Internal Use Only(非公開)

TR-I-0279

助詞の連鎖統計を用いた言語モデルとその音声認識への応用

Language Model Using N-grams for Particles and Its Application to Speech Recognition

粟津辰功† 磯谷亮輔
Shinkou AWATSU Ryosuke ISOTANI
嵯峨山茂樹
Shigeki SAGAYAMA

1992.9.4

概要

音声認識に用いられる統計的な言語モデルとして、単語の bigram,trigram などの連鎖統計が有効であることが知られている。しかし従来の方法では、文節間の係受け関係のような大域的な言語情報の表現が困難である。一方、文節ベースの連鎖統計は、そのままでは文節の種類が多いため実用的ではない。本報告では、文節内の特定の文法カテゴリに着目した単語の連鎖統計を用いることにより、より大域的な言語情報を獲得しうる言語モデルを提案する。予備的な実験として、文節末にあらわれる助詞の一文中での連鎖統計をテキストデータベースより求め、それを文節認識に適用する実験を行なったので、その結果について報告する。

†東北大学
Tohoku University
ATR 自動翻訳電話研究所
ATR Interpreting Telephony Research Laboratories

© (株)ATR 自動翻訳電話研究所 1992
© 1992 by ATR Interpreting Telephony Research Laboratories

目次

1　はじめに

音声認識に用いられる統計的な言語モデルとして、単語の連鎖統計が有効であることが知られている。しかし従来の方法では、文節間の係受け関係のような大域的な言語情報の表現が困難である。本報告では、特定の文法カテゴリに着目した単語の連鎖統計を用いることにより、より大域的な言語情報を獲得しうる言語モデルを提案する。

音声認識とは、音声波形の物理量より発声内容を推定することである。ところが物理量と発声内容とを完全に対応づけるような特徴量や手法は未だ提案されていない。そこで、入力の発声内容を推定する際に、単一の入力に対し発声内容の仮説を幾通りか立てる。これらの仮説は、同一の音声波形から推定されたものであるため音響的に似通っている。しかし、最終的にはこの複数の仮説より、最も確からしい仮説を選択しなければならない。そのため仮説を定量的に評価する必要がある。通常は入力音声と各仮説とのマッチングの良さを表す音響的尤度が用いられているが、評価をより高精度に行なうために、言語モデルが利用されている。

言語モデルとして、CFG のように文法をルールで記述する方法があるが、ルールの作成、管理が大変である。このルールに相当するものを統計量としてデータベースより自動的に獲得できる方法が望ましいと思われる。

そういった統計的言語モデルの代表的なものとして、単語ベースの bigram,trigram などの連鎖統計 (N-gram モデル) が有効であることが知られている [1]。しかし単語 bigram,trigram ではせいぜい前方 2 単語までしか参照しないため、複数文節にまたがるような表現には対応しきれなかった。

これに対して、文節 bigram,trigram が考えられるが、文節をそのまま学習したのでは、文節の種類が多過ぎるため、計算量的に困難が考えられる。

そこで、文節をそのまま学習せずに、文節内の特定の文法カテゴリの語の連鎖統計を学習することを考える。

以下に例を挙げて、説明する。

例文：「私は喉が乾いたので、水が飲みたい」

「喉」と「乾い (た)」は文中に同時に表れることが多い。しかし、単語 bigram では 1 語過去しか参照しないため「喉」と「乾い (た)」の関係を表現出来ない。また、trigram を用いたとしても、間に助詞が入るため関係を直接表現できない。

ここで、例文中の自立語のみの系列「私／喉／乾い (た) ／水／飲み (たい)」を考える。こうすると、「喉」と「乾い (た)」は bigram で表現可能になる。このように、文節内の自立語のみを利用した N-gram は、意味論的な制約を表現できる可能性がある。

付属語についても同様に考えることが出来る。

例文 1 ；「私は喉がとても乾いたので、水が飲みたい」
例文 2 ；「私は喉はとても乾いたので、水が飲みたい」

例文 1 と 2 では、1 の方が自然に思われる。これより、付属語の間にも何らかの関連性があると考えられる。つまり、「は」「は」より「は」「が」の連鎖の方が関連性が強いと考えられる。しかし単語ベースの連鎖では間にそれ以外の語があるために関係を効率的に表現できない。この場合も文節内の付属語のみを利用した N-gram を用いれば、係り受け関係のような構文的な制約を表現しうると思われる。

以上より、文節内の特定の文法カテゴリのみを利用したN-gramは、単語の共起関係のような、単語N-gramより広範囲の語の関係の表現に有効であると考えられる。しかも、この特定の文法カテゴリのN-gramはデータベースより自動的に学習可能である。

今回は、予備的な実験として、付属語のうち助詞についてN-gramの学習およびその音声認識への適用の実験・検討を行なったので報告する。

2 学習と認識のアルゴリズム

今回、学習／認識で注目する文法カテゴリは付属語のうちの助詞である。理由は次の通りである。

- 付属語の多くは助詞である。
- 活用が無く、音便が少ないので語の種類が比較的少なく、扱いやすい。
- 文節認識では、「を」を「の」や「も」と誤認識するなどの助詞の誤認識が多い。

2.1 学習

テキストデータベースより、文節の最後にくる語を取りだし、その語のうちより以下の7種類の助詞について、そのN-gramを学習する。文節末に複数の助詞が連続する場合は、簡単のため最後に表れる助詞のみに注目する。今回は認識系で品詞の情報を使用しなかったので、文法カテゴリが異なるものでも表記が同一なら、同じ語として学習している。

これ以外に、文の始端／終端を表す記号と、文節の終端の語が助詞でない場合を表す記号とを、疑似的な助詞として加える。

2.2 認識

音響的尤度に加えて、学習で得られた助詞のN-gramを言語的尤度として加味して最尤評価を行ない候補を選出する。

今回の実験では文節認識の結果に対し、後処理で言語的尤度を加味し最適文節候補列を選出する実験を行なった。

第t文節の第i位文節候補を$w_i(t)$、その音響的対数尤度$p_i(t)$とする。

音響尤度のみで最尤評価する場合は、次式を最大にする数列c_tを添字とする文節候補の系列$w_{c_t}(t)$が最適候補列となる。

音響尤度のみで最尤評価を行なうので、当然、各文節ごとに音響尤度の最も高い候補が選択される。

$$\sum_{t=start}^{end} p_{c_t}(t)$$

これに言語的尤度を加える場合は次のようになる。

- unigramの場合

音響的な尤度に文節ごとに unigram の尤度を加えて最尤評価する。$Tr_{unigram}(w_i)$ を文節 w_i の終端の助詞の unigram の尤度、$weight$ を言語的尤度の重みとして、

$$\sum_{t=start}^{end} p_{c_t}(t) + weight \cdot Tr_{unigram}(w_{c_t}(t))$$

を最大化する数列 c_t を添字とする文節候補の系列 $w_{c_t}(t)$ が最終候補の系列となる。この場合は、言語的尤度を用いない場合と同様、各文節ごとに独立に最適な候補を選択できる。

- bigram の場合

  直前の文節候補からの現在の文節候補への遷移の bigram の尤度を言語的尤度として、音響尤度に加えて最尤評価する。$Tr_{bigram}(w_i|w_j)$ を文節列 (w_j, w_i) の終端の助詞の bigram の対数尤度、$weight$ を重みとして、

  $$\sum_{t=start}^{end} (p_{c_t}(t) + weight \cdot Tr_{bigram}(w_{c_t}(t)|w_{c_{t-1}}(t-1)))$$

  を最大化する数列 c_t を添字とする文節候補の系列 $w_{c_t}(t)$ が最終候補の系列となる。

- trigram の場合

  直前の 2 つの文節候補からの現在の文節候補への遷移の trigram の尤度を言語的尤度として、音響尤度に加えて最尤評価する。$Tr_{trigram}(w_i|w_k, w_j)$ を文節列 (w_k, w_j, w_i) の終端の助詞の trigram の対数尤度、$weight$ を重みとして、

  $$\sum_{t=start}^{end} (p_{c_t}(t) + weight \cdot Tr_{trigram}(w_{c_t}(t)|w_{c_{t-2}}(t-2), w_{c_{t-1}}(t-1)))$$

  を最大化する数列 c_t を添字とする文節候補の系列 $w_{c_t}(t)$ が最終候補の系列となる。

- N-gram の場合

  以下、N-gram についても同様に考えることが出来る。

なお、bigram 以上の場合は文全体に対して最適化を行なう必要があるが、DP 法で効率的に求められる。

3 助詞の N-gram の特徴

学習に用いたテキストデータはＡＴＲ国際会議の対話データベース (12056 文 /68966 文節) である。また、学習時に注目した文法カテゴリは文節の終端に表れるすべての助詞 (表 1) であり、73 種類である。

表 1: 学習の際注目した助詞

品詞	全助詞に対する出現割合
格助詞	56.4%
接続助詞	18.7%
終助詞	10.8%
係助詞	10.2%
副助詞	2.49%
並立助詞	1.44%
準体助詞	0.043%

実際に学習した助詞の N-gram の Entropy と Perplexity を表 2 に示す。全助詞、格助詞の双方とも十分低い Entropy になっていることが分かる。つまり言語モデルとして有効であると期待できる。また、N-gram を unigram,bigram,trigram と増やしても Entropy があまり小さくならないことより、あまり N の大きい N-gram でなくとも相応の効果があると思われる。注目している情報の質が異なるため、単語の Entropy と直接の比較は出来ないが、参考に全単語、文節の最後に表れる語の Entropy と Perplexity も併記しておく。

表 2: 助詞の N-gram の Entropy と Perplexity

	格助詞 (24 種類)		全助詞 (73 種類)	
	Entropy	Perplexity	Entropy	Perplexity
unigram	1.91	3.75	3.06	8.34
bigram	1.75	3.36	2.75	6.72
trigram	1.72	3.29	2.59	6.02

	全単語 (5441 種類)		文節の終端の語 (1688 種類)	
	Entropy	Perplexity	Entropy	Perplexity
unigram	8.33	322	6.13	70.1
bigram	3.77	13.6	4.40	21.1
trigram	2.01	4.01	2.88	7.39

以下に助詞の連鎖の例を挙げる。

目立った連鎖の例

- 「から」→「へ」
- 「を」→「と」

全く見られない連鎖の例

- 「に」→「で」
- 「が」→「が」

4 認識実験

4.1 実験環境

4.1.1 音響処理部

今回、文節認識系として、SSS-LR[2][3] による不特定話者の文節認識実験を行なった。これにより、各文節に対し、第5位までの文節候補とそれぞれの音響的尤度が得られる。

そのときの、音響処理部の条件を表3、表4に示す。

表 3: 音響分析条件

分析条件	
サンプリング周波数	12kHz
ウインドウ	ハミング窓 (20ms)
分析周期	5ms
パラメータ	16 次 LPC-Cep + 16 次 Δ LPC-Cep + Power + Δ Power (34 次元)

表 4: SSS-LR による文節認識 (継続時間制御あり)

認識対象	国際会議申し込みに関する会話 7 会話 137 文 /353 文節
認識タスク (文節内文法)	ルール数 1974(単語辞書 744 語彙を含む)
話者	男性 2 名 (MAU, MNM)
パラメータ	global beam のビーム幅 256 local beam のビーム幅 64 総セル数 4000

4.1.2 言語処理部

文節終端に表れる助詞の遷移確率を言語的尤度とする。

学習に用いたテキストデータはＡＴＲ国際会議の対話データベースである。また、学習時に注目した文法カテゴリは文節の終端に表れるすべての助詞 (格助詞、接続助詞、終助詞、係助詞、副助詞、並立助詞、準体助詞) であり、全部で 73 種類である。

4.2 認識実験

4.2.1 入力フレーム長で正規化した音響尤度を用いた場合

$$(\text{尤度}) = (\text{音響的尤度}) \,/\, (\text{入力フレーム長}) + (\text{重み}) \times (\text{言語的尤度})$$

として最尤評価を行なった時の、重みと認識率の関係を表5、表6に示す。音響的尤度による候補の順位が、言語的尤度によって逆転し、正解となることが起こるため、音響的尤度のみを用いた場合 (言語的尤度の重み0) に比べ、最高、話者MAUで2.4%、話者MNMで2.6%認識率が向上している (ともに言語的尤度による認識率向上が約5%、言語的尤度により正解を不正解としてしまったものが約2.5%、実質約2.5%)。このとき音響的尤度の一位候補が正解でないもののうち約10%を修正したことになる。ただし、一位候補が正解でないものの約半数は助詞以外で間違えているので、助詞の間違いの約20%が修正されたことになる。また、重みをさらに大きくすると認識率は徐々に下がる。言語的尤度の比重が大きくなり、音響的尤度が埋もれてしまうためと思われる。

表5: 重みと文節認識率の関係 (話者MAU)

話者	MAU (音響的尤度のみでの文節認識率 77.90%)							
重み ($\times 10^{-7}$)	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
unigram	79.32%	79.32%	77.62%	75.07%	73.65%	71.95%	69.97%	69.12%
bigram	79.32%	78.75%	76.49%	75.63%	72.52%	71.38%	69.69%	68.83%

表6: 重みと文節認識率の関係 (話者MNM)

話者	MNM (音響的尤度のみでの文節認識率 88.10%)							
重み ($\times 10^{-7}$)	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
unigram	89.80%	90.65%	89.24%	87.53%	84.98%	84.13%	82.43%	81.30%
bigram	90.65%	89.52%	87.82%	86.97%	84.14%	81.30%	79.88%	78.18%

4.2.2 音響尤度に継続時間の影響を考えた場合

今回実験に用いた文節認識系SSS-LRの音響的尤度は文節全体に対する尤度を継続時間で正規化している。しかし、正規化してあるために、「△が」と「△は」との音響的尤度の差と、「□□□□□が」と「□□□□□は」との音響的尤度の差が違ったものになり、好ましくない。そこで、尤度を

$$(\text{尤度}) = (\text{音響的尤度}) + (\text{重み}) \times (\text{言語的尤度})$$

として、同じ実験を行なった。結果を表7、表8に示す。音響的尤度をそのまま用いた場合より最高性能で0.5～1%の認識率の改善が見られる。

表 7: 継続時間の影響を考えたときの重みと文節認識率の関係 (話者 MAU)

話者	MAU (音響的尤度のみでの文節認識率 77.90%)							
重み (× 0.0001)	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
unigram	78.75%	79.32%	79.32%	79.03%	78.75%	76.48%	75.92%	75.07%
bigram	79.04%	79.60%	79.32%	77.90%	76.49%	76.21%	76.21%	75.35%
trigram	79.04%	79.60%	79.32%	77.90%	76.49%	76.21%	76.21%	75.35%

表 8: 継続時間の影響を考えたときの重みと文節認識率の関係 (話者 MNM)

話者	MNM (音響的尤度のみでの文節認識率 88.10%)							
重み (× 0.0001)	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
unigram	88.67%	91.21%	90.93%	91.50%	89.80%	89.51%	88.38%	87.25%
bigram	90.08%	90.93%	91.78%	90.37%	89.80%	88.39%	87.25%	86.40%
trigram	89.74%	90.08%	90.65%	89.23%	86.12%	84.70%	83.28%	82.15%

5 まとめ

単語 N-gram より広範囲な単語間の関係を表現できる統計的言語モデルとして、特定の文法カテゴリの N-gram を提案した。予備的な実験として文節末にあらわれる助詞に注目して、テキストデータベースを用いてその N-gram を求めた。その結果日本語に特徴的な助詞の連鎖などが学習され、Entropy/Perplexity も十分低くなっていることを確認した。

また、助詞の N-gram の言語モデルによる言語的尤度を認識の後処理として用いることにより、最高で誤りの 10% を訂正することが出来、本モデルが音声認識の精度向上に有効であることが確認できた。

この言語モデルは、単語 N-gram と排他的なものではなく、単語 N-gram より上位の構造の表現に向いたものと思われる。そのかわり、特定の文法カテゴリにのみしか注目しないため、自立語と付属語の組み合わせなどの異なるカテゴリ間の連鎖関係の表現は出来ないが、これは単語 N-gram を併用することで補えると思われる。また、今回は文節認識の後処理として言語モデルを用いたが、認識アルゴリズムに直接組み込むことも可能である。

今回は時間的制約もあって、助詞についてのみの実験で終ってしまったが、この助詞の N-gram を付属語すべてに拡張し、さらに自立語の N-gram や、従来の単語 N-gram を組み合わせることにより、さらに高精度な認識を行なえる可能性があると考えている。

また、今回の実験では、文節終端には必ず助詞があるとして連鎖を学習したが、文節終端が助詞でない場合が約半数あるので、そういった文節を学習／認識の対象から外すことも考えられるが、これは今後の課題である。

謝辞

研究の機会を与えて頂いた樽松明社長と北陸先端科学技術大学院大学 下平博助教授に感謝します。また、親切に御助言、御討論下さった音声情報処理研究室の皆様に感謝します。

参考文献

[1] 南, 中川: “Trigram モデルを用いた複数候補を求める フレーム同期型 HMM 連続音声認識”, 信学論 (D), J73-D, 9, pp.1383-92(1990.9).

[2] 小坂, 鷹見, 嵯峨山: “話者混合 SSS による不特定話者音声認識と話者適応”, 音声研究会発表予定 (1992.9).

[3] 永井, 鷹見, 嵯峨山：“逐次状態分割法 (SSS) と音素コンテキスト依存 LR パーザを 統合した SSS-LR 連続音声認識システム”, 音声研資, SP92-33(1992.6).