論文 / 著書情報
Article / Book Information

論題(和文)	音声理解のための言語モデル自動獲得の検討
Title(English)	
著者(和文)	松岡 達雄, Robert Hasson, Michael Barlow, 古井 貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	1995年日本音響学会秋季講演論文集, , , pp. 21-22
Citation(English)	, , , pp. 21-22
発行日 / Pub. date	1995, 9

1—2—11

音声理解のための言語モデル自動獲得の検討*

○松岡達雄、Robert Hasson†、Michael Barlow、古井貞熙

(NTTヒューマンインタフェース研究所、†Eurecom Institute)

1. まえがき

本報告では情報検索などの音声理解システムにおいて音声認識結果（自然言語）をデータベース検索言語（意味言語）に変換するための言語モデルをコーパスから自動的に獲得する方法について述べる。Vidalらは、音声理解を自然言語から意味言語へ翻訳する問題ととらえ、自然言語と意味言語の一対となったコーパスを用いて翻訳のための言語モデルを推定する方法を提案している[1]。Vidalらの方法では自然言語の文法規則と意味言語の文法規則の共起確率を翻訳言語モデルとするが、文法規則数が多いと共起確率の推定が困難になる点が問題であった。本報告では、文法規則数が多い場合にも有効な言語モデルが推定できるようアルゴリズムの修正を行い、さらに、McCandlessらによる文脈自由文法の推定法[2]を用いて有限状態オートマトンで表現された文法の状態数を削減することにより翻訳言語モデルの推定精度を向上する方法を提案する。米国ARPAの音声理解評価タスクである航空旅行情報システム(Air Travel Information System: ATIS)を対象として評価を行い提案法の可能性を示す。

2. 音声理解システムの構成

図1に音声理解システムの構成を示す。音声理解システムは音声認識部と言語処理部からなる。

音声認識部はTree-Trellis探索により音声入力に対するN-Best仮説を生成する。N-Best探索に用いる音響モデルは単語内だけでなく単語間にわたる音素文脈も考慮している[3]。文法としては語彙単語間の任意の連接を許したno-grammarネットワークを用い、ATISドメインで学習された単語Bigramを言語モデルとして用いる。

言語処理部は音声認識結果をデータベース検索言語に変換する。ATISタスクにおけるデータベース検索言語はANSI標準のSQLであるが、ATISコーパスには一意にSQL表現に書き直せるWIN(Wizard Input)文が入力文に対応して与えられているので、WIN文を意味言語として実験を行った。

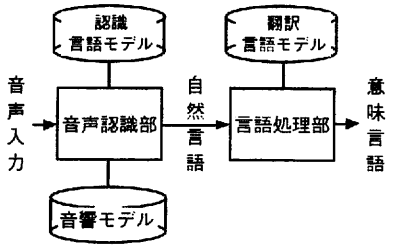

図1　音声理解システムの構成

3. 音声理解のための言語処理

3.1 翻訳言語モデルの推定

翻訳言語モデルを推定するには、まず自然言語（入力言語）、意味言語（出力言語）各々についてECGI(Errror Correcting Grammar Inference)アルゴリズム[1]により有限状態オートマトンで表現される文法を推定する。次に入力言語の文法規則と出力言語の文法規則の共起確率を入出力文が一対になった学習テキストにおける共起頻度から推定する。

入力言語の文法をG_I、出力言語の文法をG_Oとすると、与えられた問題は入力文xに対して次式を満足する出力文$\hat{y}$を求めることとなる。

$$\hat{y} = \arg\max_{y \in L(G_o)} p(y \mid x) \qquad (1a)$$

$$= \arg\max_{y \in L(G_o)} p(x \mid y) p(y) \qquad (1b)$$

文章x、yはそれぞれ文法規則の系列$D_{G_i}(x)$、$D_{G_o}(y)$として表現できる。

$$D_{G_i}(x) = \left\{ r_{I1}^x, r_{I2}^x, \ldots, r_{In}^x \mid r_{Ii}^x \in G_I \right\} \qquad (2)$$

$$D_{G_o}(y) = \left\{ r_{O1}^y, r_{O2}^y, \ldots, r_{Om}^y \mid r_{Oi}^y \in G_O \right\} \qquad (3)$$

G_I、G_Oが正規文法ならば$D_{G_i}(x)$、$D_{G_o}(y)$は一意に決定できる。文法にあいまい性がある場合にはViterbiアルゴリズムによって近似的に求めることができる。これより、(1)式は(4)式のように書き直せる。

$$\hat{y} = \arg\max_{y \in L(G_o)} p\left(D_{G_o}(y) \mid D_{G_i}(x)\right) \qquad (4a)$$

$$= \arg\max_{y \in L(G_o)} p\left(D_{G_i}(x) \mid D_{G_o}(y)\right) p(y) \qquad (4b)$$

Vidalらは、

$$p\left(D_{G_i}(x) \mid D_{G_o}(y)\right) \approx \prod_{r_O \in D_{G_O}(y)} \left(\prod_{r_I \in D_{G_I}(x)} p(r_I \mid r_O) \prod_{r_I \notin D_{G_I}(x)} p(not\ r_I \mid r_O) \right) \qquad (5)$$

として(4b)式に基づく推定を試みている。(5)式では入力文法規則r_Iが出力文法規則r_Oと共起しない確率$p(not\ r_I \mid r_O)$を用いている。文法規模が非常に小さい場合には$p(not\ r_I \mid r_O)$が有効な場合も考えられるが、現実的なタスクを扱う文法の場合には、本質的に無関係な文法規則間の共起頻度を有意な共起頻度と同等に評価することによる悪影響がある

---

* Language model acquisition from a text corpus for speech understanding,
Tatsuo Matsuoka, Robert Hasson, Michael Barlow, and Sadaoki Furui
(NTT Human Interface Laboratories, Eurecom Institute)

と考えられる。計算上も項数の増加により確率の積が非常に小さくなるなどの悪影響がある。我々は、出力文法の方が入力文法より規模が小さいことより、(4a)式に基づいて、$p\left(D_{G_i}(x)\mid D_{G_o}(y)\right)$ではなく$p\left(D_{G_o}(y)\mid D_{G_i}(x)\right)$を推定する方が有利と考え、

$$p\left(D_{G_o}(y)\mid D_{G_i}(x)\right)\approx p\left(r_O\mid r_{I1}^x,r_{I2}^x,\ldots,r_{In}^x\right)\approx\frac{N(r_O,x)}{N(x)}\qquad(6)$$

として推定することとした。$N(x)$はG_Iが生成可能な文全体の数であり実際には計算不可能なので以下のように近似を行う。

$$\hat{p}\left(r_O\mid r_{I1}^x,r_{I2}^x,\ldots,r_{In}^x\right)\approx\frac{p^{\alpha}(r_O)\prod_{r_I\in D_{G_I}(x)}p^{\beta}(r_O\mid r_I)}{\sum_{r\ with\ the\ same\ initial\ state\ as\ r_O}\left[p^{\alpha}(r)\prod_{r_I\in D_{G_I}(x)}p^{\beta}(r\mid r_I)\right]}\qquad(7)$$

以上に述べた翻訳言語モデルの推定では、文法G_I、G_Oの規模が大きいほど、つまり、文法規則数が多いほど$p\left(D_{G_o}(y)\mid D_{G_i}(x)\right)$の推定がスパースになるという問題がある。したがって文法規則数が少ないことが望ましい。

ECGIアルゴリズムによって推定される文法の状態数は学習セット中の文の提示順序により異なる。すなわち、単語数の多い文から提示すれば、単語数の少ない文はすでに有限状態オートマトンに存在する状態間を遷移することになり、結果として2割ほど状態数の少ない文法が推定される。しかし、さらに文法規模を縮小するため次節に述べるような方法で有限状態オートマトンの状態数を削減した。

3.2 文脈自由文法の推定による文法状態数の削減

McCandlessらはコーパスから確率的文脈自由文法を推定する方法を提案している[2]。この方法では単語Bigram確率を距離尺度として、ボトムアップに文法規則を生成する。単語や非終端記号間の距離は(8)～(10)式のように定義する。

$$\left\|u_i,u_j\right\|=d(P_i,P_j)+d(P_j,P_i)\qquad(8)$$

$$d(P_i,P_j)=\sum_{C\in Context}P_i(C)\times\log\frac{P_i(C)}{P_j(C)}\qquad(9)$$

$$P_i(C)=P(C\mid u_i)\approx P(u_{left},u_i\mid u_i)\times P(u_i,u_{right}\mid u_i)\approx\frac{N(u_{left},u_i)}{N(u_i)}\times\frac{N(u_i,u_{right})}{N(u_i)}\qquad(10)$$

すなわち、同じ文脈で出現する頻度の高い単語同士の距離を近いと判断しマージして新たな非終端記号を生成する。単語がマージされて非終端記号となることで有限状態オートマトンの状態数を削減することができる。

3.3 前処理

タスク固有の知識を必要としない前処理によっても文法上の状態数の削減が可能である。例えば、Boston、New Yorkなどの単語は「地名」という一つの非終端記号にまとめることができる。同様に、数字、日付、曜日、月、航空会社、空港名、座席クラスなども具体的な単語の代わりに非終端記号で置き換え可能である。文脈自由文法による非終端記号へのマージと、この前処理により入力言語の文法状態数は約2500から約1500へ、出力言語の状態数は約1000から約600へ削減できた。

4. 実験

LDC(Linguistic Data Consortium)より頒布されているATIS2のコーパスより文脈に独立にWIN文を生成可能なクラスAに分類されたデータのうちWIN文に括弧の含まれないデータを対象として実験を行った。1915文を学習に、213文を評価に用いた。

ECGIアルゴリズムによって推定された文法の評価セットに対するカバー率は約90%であった。また、同じ非終端記号が有限状態オートマトンの異なる場所にいくつも存在するなど文法に冗長性があることがわかった。

評価は翻訳結果と正解WIN文との文章単位での正解率により行った。(7)式のβは1-αとして、αを実験的に最適化して実験を行った結果、学習セットに対する文正解率は95.3%、評価セットに対しては62.4%であった。評価セットに対する性能はまだ十分とは言えない。しかし、この結果はコーパスからの自動推定を重視してタスク固有のヒューリスティックスなどは用いずに評価しているため、ARPAにおける他の評価結果と直接比較することはあまり意味がない。

5. まとめ

音声理解のために自然言語を意味言語に翻訳する言語モデルをコーパスから自動的に推定する方法について述べた。Vidalらの方法において文法の規模が大きくなった場合の問題点を明らかにし、アルゴリズムを修正するとともに、文法上の状態数を削減して翻訳言語モデルの推定精度を向上する方法を提案した。ATIS2のコーパスにより評価を行いコーパスから翻訳言語モデルを自動獲得することの可能性を示した。ECGIアルゴリズムにより推定される文法の冗長性や、文脈自由文法の推定において同一コンテキストでの出現頻度のみで単語をマージしていることなどが問題として考えられる。今後これらの問題点を解決するとともに、タスク固有のヒューリスティックスも導入容易なものについては考慮しながら、音声入力から意味理解までの全体的な評価を行っていきたい。

謝辞

ECGIアルゴリズムのプログラムを提供してくださったUniv. Politecnica de ValenciaのVidal教授と、修士論文を送ってくださったMITのMcCandless氏に感謝します。

参考文献

1. E. Vidal, R. Pieraccini, and E. Levin, "Learning associations between grammars: a new approach to natural language understanding," Proc. EUROSPEECH-93, pp. 1187-1190
2. M. K. McCandless and J. Glass, "Empirical acquisition of word and phrase classes in the ATIS domain," Proc. EUROSPEECH-93, pp. 981-984
3. W. Chou, T. Matsuoka, B. H. Juang, and C. H. Lee, "An algorithm of high resolution and efficient multiple string hypothetization for continuous speech recognition using inter-word models," Proc. ICASSP-94, pp.II-153-156