音声言語情報処理 14－15
(1996. 12. 13)

統計的翻訳言語モデルを用いた音声理解

松岡達雄、古井貞熙

NTTヒューマンインタフェース研究所

〒180 東京都武蔵野市緑町3-9-11
matsuoka@splab.hil.ntt.jp

あらまし 本報告では、コーパスから自動的に学習した翻訳言語モデルを用いて、音声認識結果である自然言語をシステムを駆動する意味言語に変換する音声理解の方法について述べる。翻訳言語モデルの学習では、まず、自然言語／意味言語における単語を、出現する文脈の類似度を尺度とした統計的なクラスタリングによりグループ化する。次に、自然言語、意味言語をそれぞれネットワーク文法で表現し、自然言語の文法ネットワーク中の状態遷移と、対応する意味言語の文法ネットワーク中の状態遷移間の共起確率を、自然言語と意味言語が一対となったコーパスを用いて推定する。この共起確率を翻訳言語モデルとして、自然言語から意味言語への変換を行う。単語のクラスタリングによりネットワーク中の状態数が削減されているため、スパースデータからの推定の問題を回避し、頑健な翻訳言語モデルを推定することができる。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、提案法の有効性を示す。
キーワード 音声理解、翻訳、言語モデル、自然言語、意味言語

Speech understanding using a statistical translation language model

Tatsuo Matsuoka and Sadaoki Furui

NTT Human Interface Laboratories

3-9-11 Midori-cho, Musashino-shi, Tokyo 180
matsuoka@splab.hil.ntt.jp

Abstract This paper describes a speech understanding method that uses a translation language model estimated automatically from a text corpus. The translation language model translates the natural language output by a speech recognition system into semantic language. For training the translation language model, words in natural and semantic languages are first clustered using a measure of word contextual similality. Natural and semantic languages are then modeled using grammar networks with the word clusters as nodes in the networks. Co-ocurrence probabilities of transitions in natural-language and semantic-language grammar networks are estimated as parameters of the translation language model. This method was shown to be very effective by experiments using the ARPA ATIS speech understanding evaluation task.
Keywords Speech understanding, Translation, Language modeling, Natural language, Semantic language

1. まえがき

我々は米国ARPA(Advanced Research Projects Agency)の音声理解標準評価タスクであるAir Travel Information System（航空旅行情報案内システム、以下ATIS）[1]を対象タスクとして、音声理解の研究を進めている[2-7]。音声理解を実現するためには、大きく分けて音声入力を自然言語である単語列に変換する音声認識の機能と、単語列から意味を抽出する言語処理の機能が必要である。音声認識に関しては、我々は、これまでに、音素コンテキストを考慮した詳細な音響モデルを用いたN-best探索法による音声認識システムを構築した[8]。言語処理に関しては、音声認識結果の自然言語をデータベース検索言語である意味言語に変換（翻訳）する翻訳言語モデルを、コーパスから自動的に構築する方法を提案した[3-7]。本報告では、N-best探索による音声認識結果を用いて、音声理解システムの言語処理部の評価を行った結果を述べる。

音声理解のための言語処理は、これまで主に人手により文法規則を書いていく方法で実現されていたが、文法の構築に専門的知識が必要であること、開発に要する時間・労力の問題、異なるタスクへの移植性の問題などから、最近、自動的に文法規則あるいは言語モデルを構築する方法の研究が盛んになってきている[9-17]。Kuhnらは、木構造を用いた単語のクラス分けによる、コーパスからの文法の獲得法を提案しているが、これまでのところあまり高い性能は得られていない[9]。Unisysは、言語処理部を特定タスクへ適応化する方法として、出力が正解であるか否かだけを教師信号として与える弱教師あり学習法を提案しているが、あまり大きな効果は得られていない[11]。また、人間が正解を与えるという意味で自動的な方法ではない。BBNは自然言語と意味言語を木構造の言語モデルで表現し、その対応関係をHidden Understanding Model (HUM)という統計的言語モデルによりモデル化し、コーパスから学習する方法を提案している[12, 13]。しかし、木構造の言語モデルを生成するためにコーパスに対してタグ付け(annotation)が必要である。また、木構造が意味構造を反映したものであるため、意味抽出のための言語モデルはコーパスのタグ付けの方法（タグの種類など）に依存する[12]。また、BBNの音声理解システムにおける言語処理は、文→初期意味構造→正規化意味構造→意味、と三段階にモジュール化されているが、第一段階の処理にHUMを用い、残りの処理は人手で書かれたルールベースの言語処理によって実現されている[13]。AT&Tは、言語処理のための知識源のうち一般的な知識源とタスク固有の知識源の分離を図り、また、データから学習可能な部分については学習することで移植性の向上を図っている[14]。

AT&Tのシステムの中心をなす概念的デコーダ(Conceptual Decoder)は、Pieraccinらにより提案された意味構造を統計モデルを用いて学習／抽出する方法である[15, 16]。この方法では、音声理解の問題を

$$P(\hat{W},\hat{C}\mid A)=\max_{W,C}P(W,C\mid A) \qquad (1)$$

ただし、A:観測音響信号、W:単語列、C:概念ラベル列を満たす$\hat{C}$を求める問題として定式化する。ここで、

$$P(W,C\mid A)=\frac{P(A\mid W,C)P(W\mid C)P(C)}{P(A)} \qquad (2)$$

であるから、$P(A\mid W,C)\approx P(A\mid W)$を音響モデル、$P(W\mid C)P(C)$を概念言語モデルとしてそれぞれHMMとして推定する。しかし、$P(W\mid C)P(C)$の推定には概念ラベル列が必要であり、コーパスのタグ付け(annotation)が必要となる。また、BBNのHUMと同様に、概念言語モデルはタグ付けの方法に依存したものとなる。

Vidalらは、Pieraccinの枠組みにおける概念を文法規則（文法ネットワークにおける状態遷移）に置き換えた方法を提案している[17]。この方法の利点は、コーパスへのタグ付けが必要なく、また文法ネットワークの生成もデータから自動的に行なわれるため、コーパスさえあれば言語処理に関する専門知識を一切必要とせず、自動的に言語処理部を構築できる点である。この方法は、Brownらの統計的機械翻訳の方法[18, 19]にヒントを得ており、Brownらの方法における語順の問題を解決した方法である。Brownらの方法は、自然言語から自然言語への翻訳を対象として、入力言語、出力言語とも単語n-gramによりモデル化していた。このため、語順が直接的に統計量に反映されてしまう。しかし、自然言語から意味言語への翻訳では、同一の意味の発話は自然言語の語順に依存せずに同一の意味言語に翻訳されなければならない。Vidalらの方法ではまず自然言語、意味言語の文法ネットワークを学習テキストから推定し、それら文法ネットワーク上の状態遷移（文法規則）間の条件付き確率を翻訳言語モデルとする。Vidalらは、ATISコーパスのうち、文の構造が簡単なサブセットを選択して評価を行い、90%程度の文が正確に意味言語に翻訳できたと報告している。しかし、対象データを、例えばATISのクラスA（文脈独立、単文で回答可能）全体に広げると、推定すべき翻訳言語モデルのパラメータ空間が非常に大きくなり、通常の計算機環境では計算不可能、かつ、限られたデータからは精度のよい推定ができないという問題点がある。現実には、自然言語の表現のバリエーションは大きく、文法規則数は多くな

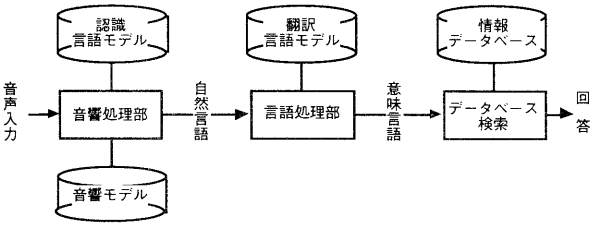

図1　音声理解システムの構成

る。文献[17]における翻訳言語モデルの定式化では、文法規則数が多い場合に確率計算におけるアンダーフローにより計算不可能となりやすいという問題点もある。

我々は、入力される自然言語のバリエーションが大きい場合にも、翻訳言語モデルの推定を可能とする方法を提案した[3-7]。提案法は、単語（および単語列）間の文脈の類似度を距離尺度として統計的なクラスタリングを行い、単語（および単語列）を単語クラスに置き換えることで、自然言語、意味言語について有限状態オートマトンで表現された文法ネットワークの状態数を削減することにより、より広範な表現に対しても翻訳言語モデルの推定を可能とし、また、文献[17]の翻訳言語モデルの定式化を見直し、文法規模が大きい場合にも推定の問題を生ずることのないような定式化に修正したものである。本報告では、ATISタスクを対象として、音声認識結果を用いた自然言語処理の評価を行い提案法の有効性を示す。

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[8]。音声認識のための文法としては、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。

3. 統計的翻訳言語モデル

図2に統計的言語モデルを用いた音声理解のための言語処理の流れを示す。上段は翻訳言語モデルを推定する過程、下段はその翻訳言語モデルを用いて自然言語の文を意味言語の文に翻訳する過程を示している。翻訳言語モデルの推定では、まず、入力自然言語、出力意味言語のそれぞれの学習テキストセットに対して、文法ネットワークの状態数を少なくするため前処理を行う。前処理により単語を単語クラスに変換した学習テキストセットを用いて文法ネットワークを生成する。自然言語文と意味言語文の各対について文法規則列（すなわち、状態遷移系列）を求め文法規則が適

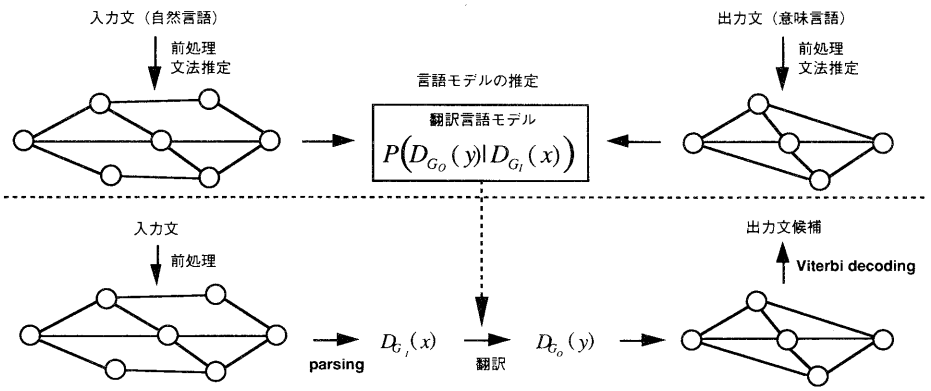

図2　統計的言語モデルによる音声理解の言語処理

用される条件付き確率を推定する。下段の翻訳処理においては、入力文にまず前処理を施し、文法ネットワークによりパージングして得られる文法規則系列$D_{G_I}(x)$に対して、先に求められた文法規則の条件付き確率$P\left(D_{G_O}(y)\mid D_{G_I}(x)\right)$を翻訳言語モデルとして適用し、出力側の文法規則系列$D_{G_O}(y)$を求める。この$D_{G_O}(y)$に対応する最適パスをViterbi探索により出力文法ネットワーク内で求めれば、出力文候補を決定できる。最後に単語クラスに置き換えられた単語は、もとの単語に戻す。これは、前処理の履歴を保持していれば可能である。

3.1 前処理

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を少なくすることが重要である。文法規模を縮小し、精度のよい翻訳言語モデルを推定するため、以下のような方法で文法ネットワークの状態数を削減した。

(1) 一般的知識による前処理

例えば、Boston、New Yorkなどの単語は、特別な専門知識がなくとも「地名」という単語クラスにまとめることができる。数字、日付、曜日、月、都市名、航空会社、空港名、座席クラスなどは、ごく一般的な知識により単語クラスにまとめられるので、これらについては、次節に述べる単語クラスタリングの収束を速めるために、クラスタリングを行う前に単語クラスにまとめた。

(2) 単語クラスタリングによる文法状態数の削減

文法ネットワーク上の各状態は単語であるから、単語（あるいは単語列、以下同様に単語は単語列も含む）を何らかの基準で単語クラスにまとめることができれば、文法ネットワークの状態数を削減することが可能であり、推定すべきパラメータ空間を小さくすることができる。音声理解の観点からは、単語を意味に直結するようなクラスに分類できれば、非常に効率のよい翻訳言語モデルを推定できると期待できる。しかし、そのためにコーパスにタグ付けすることは、専門的知識を用いず、コーパスから自動的に翻訳言語モデルを推定したいとする我々の目的に反する。そこで、単語の文脈を考慮した統計的クラスタリングにより単語を分類することを考えた。クラスタリングの距離尺度としては、単語bigram確率のdivergenceを用いる。この距離尺度は、E. Brillらが、コーパスからの句構造の自動獲得[20]に、McCandlessらが、コーパスから確率的文脈自由文法を推定する方法[21, 22]に用いて良好な結果を得ている。同一文脈での単語出現頻度は、意味と直接的に関係していないが、似たような概念の単語は高い頻度で同一文脈に出現すると期待できるので、ある程度、意味との関連性はあると考えられる。また、クラスタリングの結果、意味的に一つのクラスになっては都合の悪い単語クラスは、採用しないことも可能である。

距離尺度の定義は(3)～(5)式の通りである。

$$\left\|u_i,u_j\right\| = d(P_i,P_j)+d(P_j,P_i) \tag{3}$$

$$d(P_i,P_j) = \sum_{C\in Context} P_i(C)\times\log\frac{P_i(C)}{P_j(C)} \tag{4}$$

$$\begin{aligned} P_i(C) &= P(C\mid u_i)\\ &\approx P(u_{left},u_i\mid u_i)\times P(u_i,u_{right}\mid u_i)\\ &\approx \frac{N(u_{left},u_i)}{N(u_i)}\times\frac{N(u_i,u_{right})}{N(u_i)} \end{aligned} \tag{5}$$

uは単語あるいは単語クラス、Cは文脈、$N(u)$はuの観測回数である。同じ文脈で出現する頻度の高い単語同士をマージして新たな単語クラスを生成する。単語がマージされて単語クラスとなることで、文法ネットワークの状態数を削減することができる。

表1　単語クラスタリングによる文法状態数削減の効果

	自然言語の文法状態数	意味言語の文法状態数	文法状態数の積	翻訳言語モデルの容量	翻訳結果（文誤り率）
前処理なし	3453	1473	5.1 M	計算不可能	-
一般的知識による前処理	2179	1070	2.3 M	191 MB	56.3 %
単語クラスタリング	1392	637	0.9 M	87 MB	25.4 %

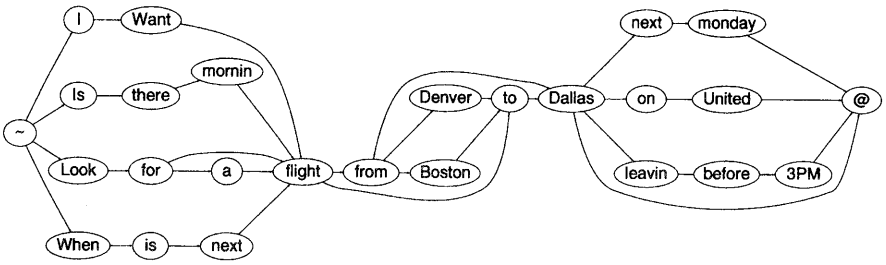

図3　ECGIアリゴリズムにより生成される文法ネットワークの例

前節で述べた一般的知識に基づく前処理と、単語クラスタリングによる前処理による、文法ネットワークの状態数削減の効果を表1に示す。翻訳言語モデルの推定すべきパラメータ数は自然言語、意味言語の各ネットワークの状態数の積にほぼ比例するので、文法状態数の積を一緒に示した。前処理なしの場合、約400MBのメモリを実装したUltra SPARCではメモリ不足により翻訳言語モデルは推定不可能であった。単語クラスタリングによって、一般的知識を用いた前処理より、さらに翻訳言語モデルの容量を1/2以下に削減できた。4章に述べる実験条件のもとで、テキストを入力した場合の自然言語から意味言語への翻訳の文誤り率（全単語完全一致で正解）も1/2以下にできた。

3.2　文法ネットワークの生成

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）各々について有限状態オートマトンで表現される文法ネットワークを生成する。文法ネットワークはどのような方法で生成しても構わないが、我々は、コーパスからの翻訳言語モデルの自動獲得を目的としているため、テキストデータから自動的に文法ネットワークを生成できる方法として、Error Correcting Grammar Inference(ECGI)アルゴリズム[23]を用いた。ここで、意味言語に関してもコーパスからの自動推定を行ったのは、意味言語として用いるWIN文の表現が収録機関や収録時期の違いにより異なり、必ずしも一貫していないためである。また、人手で、約3000文の学習テキストを見ながら、全ての表現を網羅するような文法を書き下すより、コーパスから自動生成する方が圧倒的に短時間で文法ネットワークを構築できる。（文法ネットワーク推定の計算時間は、2781文を学習に用いた場合、Ultra SPARC上で数分程度である。）

ECGIアルゴリズムは、学習セット中のテキストを一文ずつパージングし、その文を受理するために必要な状態、遷移を付加していく。入力文と文法ネットワーク中のパスの最適アライメントをDP的手法により求め、その最適パスに沿って必要な状態、遷移を付加する。入力文Xと文法ネットワークにより生成される文Yとの最適アライメントを求めるための距離尺度には、次式で定義されるLevensteinの距離を用いた。

$$D(X,Y)=\min_{s}(p\cdot sub_s+q\cdot ins_s+r\cdot del_s) \qquad (6)$$

ここで、p、q、rは重み係数で、sは文法ネットワークにおける状態系列、sub_sは置換誤り、ins_sは挿入誤り、del_sは脱落誤りである。実験では$p=q=r=1$とした。図3は、ECGIアルゴリズムにより生成される文法ネットワークの例である。

実験から、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ことがわかった。これは、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ネットワーク上に存在する状態間を遷移することになり、その結果、状態数の少ない文法が生成されるためと考えられる。

文の提示順序を、単語数で数えて昇順、降順、コーパス中にあった通りの順（ソートなし）とした場合に生成される文法ネットワークの状態数がどのようになるか検討した結果を表2に示す。降順で提示したものは、コーパス中にあった順で提示するのに比較して36%文法状態数の積（パラメータ空間）が小さくできている。昇順と降順の場合では3倍以上の差があることから、文の提示順序を考慮することが重要である。

一方で、この結果は、アルゴリズムの評価基準が、学習セット中のすべての文を同時に評価するグ

表2　文の提示順序による状態数の変化

	自然言語の文法状態数	意味言語の文法状態数	文法状態数の積
昇順	2877	1112	3.2 M
ソートなし	1815	764	1.4 M
降順	1392	637	0.9 M

ローバルな尺度になっていないことを示している。詳しく調べると、同じシンボル（単語／単語クラス）が、生成された文法ネットワークの異なる場所に複数存在するなど、多少冗長性があることがわかった。この部分については、今後マルコフモデル、HMMなど学習セット全体を同時に評価する基準で文法ネットワークの推定をする方法にしていく必要があると考えられる。

3.3 翻訳言語モデルの推定

次に入力言語の文法規則と出力言語の文法規則の条件付き確率を、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。

$$\hat{y} = \arg\max_{y \in L(G_o)} P(y \mid x) \tag{7}$$

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

$$D_{G_I}(x) = \left\{ r_{I1}^x, r_{I2}^x, ..., r_{In}^x \mid r_{Ii}^x \in G_I \right\} \tag{8}$$

$$D_{G_o}(y) = \left\{ r_{O1}^y, r_{O2}^y, ..., r_{Om}^y \mid r_{Oi}^y \in G_O \right\} \tag{9}$$

G_I、G_Oが正規文法ならば、、$D_{G_o}(y)$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、(7)式は(10)式のように書き直せる。

$$\begin{aligned} \hat{y} &= \arg\max_{y \in L(G_o)} P\left(D_{G_o}(y) \mid D_{G_I}(x)\right) \\ &\approx \arg\max_{y \in L(G_o)} P\left(r_O \mid r_{I1}^x, r_{I2}^x, ..., r_{In}^x\right) \\ &\approx \arg\max_{y \in L(G_o)} \frac{N(r_O, x)}{N(x)} \end{aligned} \tag{10}$$

$N(x)$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。

$$\begin{aligned} &\hat{P}\left(r_O \mid r_{I1}^x, r_{I2}^x, ..., r_{In}^x\right) \\ &\approx \frac{P^\alpha(r_O) \prod_{r_I \in D_{G_I}(x)} P^\beta(r_O \mid r_I)}{\sum_{r \text{ with the same initial state as } r_O} \left[P^\alpha(r) \prod_{r_I \in D_{G_I}(x)} P^\beta(r \mid r_I) \right]} \end{aligned} \tag{11}$$

4. 実験

LDC(Linguistic Data Consortium)より頒布されているATIS0/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。ARPAのコンテストにおいて学習セットに指定されているデータに加えて、FEB92テストセットを学習に用い、NOV92テストセットを評価セットとした。学習セットは2781文、評価セットは279文からなる。

4.1音声認識実験

音声認識には文献[8]の音声認識システムを用いて、10位までのN-best仮説を求めた。NOV92テストセットに対する音声認識結果を表3に示す。この結果は、クラスD（文脈依存な文、単文で回答不可能）、クラスX（回答不可能）も含んでいる。収録機関ごとに、シナリオが異なるためか、発話の表現に差が大きく、特にAT&T収録の音声はspontenuityが高いためか、単語誤り率が高い。全体での単語誤り率は、1位仮説で14.3%、10位仮説までを考慮した場合で9.7%であった。

4.2 自然言語から意味言語への翻訳

翻訳精度を文誤り率で評価した結果を、表4、表5に示す。(11)式のα、βは収録機関ごとに実験的に最適化した。表4は、翻訳結果と正解WIN文とが全単語完全一致した文を正解とした場合である。この場合、学習セットに対するテキスト入力の場合の文誤り率は16.8%であった。

表5は意味的に正しいものを正解とした場合で、例えば、

翻訳結果：List late evening filghts from cityA to cityB having prices of fares associated with economy class service

正解WIN：List late evening flights from cityA to cityB having prices of fares associated with fare bases whose

表3 音声認識結果

収録機関	単語誤り率	
	1位	10 best
AT&T	22.9 %	18.8 %
BBN	15.5 %	10.0 %
CMU	10.7 %	6.2 %
MIT	12.3 %	7.4 %
SRI	11.0 %	6.8 %
Total	14.3 %	9.7 %

表4　文誤り率で評価した翻訳結果
（完全一致）

収録機関	言語処理への入力データ		
	テキスト	音声認識結果	
		1位	10 best
AT&T	17.4 %	26.1 %	26.1 %
BBN	30.1 %	49.3 %	39.7 %
CMU	12.8 %	21.3 %	14.9 %
MIT	33.3 %	40.9 %	39.4 %
SRI	24.3 %	28.6 %	27.1 %
Total	25.4 %	35.5 %	31.2 %

表5　文誤り率で評価した翻訳結果
（意味的に正解）

収録機関	言語処理への入力データ		
	テキスト	音声認識結果	
		1位	10 best
AT&T	17.4 %	21.7 %	21.7 %
BBN	21.9 %	39.7 %	28.8 %
CMU	10.6 %	21.3 %	12.8 %
MIT	19.7 %	24.2 %	22.7 %
SRI	18.6 %	22.9 %	22.9 %
Total	18.6 %	26.9 %	22.6 %

economy is yes

のように、表記上の違いがあっても、意味的に正しいものは正解とした。音声認識誤りのないテキストを入力した場合には、文誤り率は18.6%、音声認識結果を用いた場合にも、10位までの認識結果を考慮することで、22.6%の文誤り率となっている。

NOV92テストセットのクラスAに対する、ARPAにおける各研究機関の評価結果(NL test results、unweighted error rate)は12.2%～79.9%に分布している[24]。この結果は、ARPAのコンテストのために規定されたフォーマットおよびプログラムによる単語誤り率での評価であるため、直接比較はできないが、表5の結果を単語誤り率になおすと、26.9%(1位)は13.6%、22.6%(10 best)は7.2%となる。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を用いていないことを考慮すれば、非常に有望な結果と考えられる。

本実験で対象とした文（発話）は、極めてspontaneousで、言い淀み、言い直し、感嘆詞、言葉でない音声などを含んでいる。提案法は、単語を上位の単語クラスに変換し、また、コーパスから統計的に翻訳言語モデルを推定するため、たとえ文法的に正しくなくとも、頑健に意味言語に翻訳することができる。例えば、評価セット中にあった以下のような発話から完全一致で正しく意味言語が生成できている。

"I would like do you have any flights between Philadelphia and Atlanta。"

"Oh dear I would prefer to fly first class on American from Philadelphia to Dallas and I would like to leave in morning I would like to know if there is any such flight。"

最初の例では、I would likeやdo you haveは、show meやtell meなどとともに、要求を意味するような単語クラスにまとめられるため、矛盾なく翻訳できる。二番目の例では、Oh dearはOkayやThank youなどとともに意味言語への翻訳にとっては特に意味を持たない単語クラスとなるため翻訳に関して問題を起こさない。

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。単語（および単語列）の文脈を考慮した統計的クラスタリングにより単語（および単語列）を単語クラスに分類し、自然言語の表現のバリエーションが大きい場合にも、効率よく頑健な翻訳言語モデルを推定できる方法を提案した。ATISタスクにより評価を行い、コーパスから自動獲得した翻訳言語モデルの有効性を示した。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文脈での出現頻度のみで単語をマージしていることなどは、今後の課題として改善の余地がある。今後これらの課題を検討するとともに、導入が容易なものについては、タスク固有のヒューリスティックスも考慮し、さらに性能向上を図りたい。

謝辞

ECGIアルゴリズムのプログラムを提供して下さったUniversidad Politécnica de ValenciaのEnrique Vidal教授に感謝します。統計的機械翻訳の関連研究をご教示下さった東京工業大学田中穂積教授、統計的機械翻訳の関連論文を送って下さったNTT情報通信研究所永田昌明氏、修士論文を送って下さったMITのMichael McCandless氏に感謝します。

参考文献

1. MADCOW、 "Multi-Site Data Collection for a Spoken Language Corpus、" Proc. DARPA Speech and Natural Language Workshop、 pp. 7-14、 Feb. 1992

2. M. Barlow、 T. Matsuoka、 and S. Furui、 "Markov model recordering of sentence hypothses、" 日本音響学会平成7年度春季研究発表会、3-P-7、pp. 175-176、Mar. 1995
3. 松岡達雄、R. Hasson、M. Barlow、古井貞熙、" 音声理解のための言語モデル自動獲得の検討、" 日本音響学会平成7年度秋季研究発表会、1-2-11、pp. 21-22、Sep. 1995
4. 松岡達雄、R. Hasson、M. Barlow、古井貞熙、" テキストコーパスを用いた音声理解のための言語モデル自動獲得、" 信学技報、NLC95-59、SP95-94、pp. 7-12、Dec. 1995
5. 松岡達雄、R. Hasson、M. Barlow、古井貞熙、" 音声理解のための言語モデル自動獲得、" 1996信学総全大、情報・システム1、pp. 347-348、Mar. 1996
6. T. Matsuoka、 R. Hasson、 M. Barlow、 and S. Furui、 "Language model acquisition from a text corpus for speech understanding、" Proc. ICASSP-96、 pp. I-413-416、 May 1996
7. 松岡達雄、R. Hasson、S. Dal、M. Barlow、古井貞熙、" テキストコーパスを用いた音声理解のための言語モデル自動獲得、" 信学論 D-II、Vol. J79-D-II、No. 12、Dec. 1996
8. 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、 Apr. 1994
9. R. Kuhn and R. De Mori、 "Learning speech semantics with keyword classification trees、" Proc. ICASSP-93、 pp. II-55-58、 Apr. 1993
10. D. A. Dahl、 "Summary of ATIS sessions、" Proc. ARPA Spoken Language Technology Workshop、 pp. 241-242、 Jan. 1995
11. D. A. Dahl、 L. M. Norton、 C. E. Weir、 and M. C. Linebarger、 "Weakly supervised training for spoken language understanding systems、" ARPA Spoken Language Technology Workshop、 pp. 272-275、 Jan. 1995
12. S. Miller、 R. Bobrow、 R. Ingria、 and R. Schwartz、 "Hidden understanding models of natural language、" Proc. the 32nd Annual Meeting of the Association for Computational Linguistics、 pp. 25-32、 1994
13. S. Miller、 M. Bates、 R. Bobrow、 R. Ingria、 J. Makhoul、 and R. Schwartz、 "Recent progress in hidden understanding models、" Proc. ARPA Spoken Language Technology Workshop、 pp. 276-280、 Jan. 1995
14. E. Levin and R. Pieraccini、 "CHRONUS、 The next generation、" Proc. ARPA Spoken Language Technology Workshop、 pp. 269-271、 Jan. 1995
15. R. Pieraccini and E. Levin、 "Stochastic Representation of Semantic Structure for Speech Understanding、" Proc. EUROSPEECH-91、 Proc. Vol. 2、 pp. 383-386、 1991
16. R. Pieraccini、 E. Levin、 and E. Vidal、 "Learning how to understand language、" Proc. EUROSPEECH-93、 Vol. 2、 pp. 1407-1412、 Sep. 1993
17. E. Vidal、 R. Pieraccini、 and E. Levin、 "Learning associations between grammars: a new approach to natural language understanding、" Proc. EUROSPEECH-93、 pp. 1187-1190、 Sep. 1993
18. P. F. Brown、 J. Cocke、 S. A. D. Pietra、 V. J. D. Pietra、 F. Jelinek、 J. D. Lafferty、 R. L. Mercer、 and P. S. Roossin、 "A statistical approach to machine translation、" Computational Linguistics、 Vol. 16、 No. 2、 pp. 79-85、 1990
19. P. F. Brown、 S. A. D. Pietra、 V. J. D. Pietra、 J. D. Lafferty、 and R. M. Mercer、 "Analysis、 Statistical Transfer、 and Synthesis in Machine Translation、" Proc. International Conference on Theoretical Methodological Issues in Machine Translation、 pp. 83-100、 1992
20. E. Brill and M. Marcus、 "Automatically acquiring phrase structure using distributional analysis、" Proc. DARPA Speech and Natural Language Workshop、 pp. 155-159、 Feb. 1992
21. M. K. McCandless and J. Glass、 "Empirical acquisition of word and phrase classes in the ATIS domain、" Proc. EUROSPEECH-93、 pp. 981-984、 Sep. 1993
22. M. K. McCandless、 "Automatic Acquisition of Language Models for Speech Recognition、" Thesis for M.E.、 Dept. of Electrical Engineering and Computer Science、 MIT、 Jun. 1994
23. H. Rulot、 N. Pietro、 and E. Vidal、 "Learning Accurate Finite-State Structual Models of Words through the ECGI Algorithm、" Proc. ICASSP-89、 pp. 643-646、 May 1989
24. D. S. Pallett、 J. G. Fiscus、 W. M. Fisher、 and J. S. Garofolo、 "Benchmark tests for the DARPA spoken language program、" Proc. ARPA Human Language Technology Workshop、 Mar. 1993