確率モデルによる発話の最適分割と意図認識

田中 英輝　　横尾 昭男
ATR 音声翻訳通信研究所
{tanakah, ayokoo}@itl.atr.co.jp

1 はじめに

本稿は音声翻訳システムにおける発話意図の自動認識手法について述べる。これは談話タグ付与問題の一種、すなわち談話単位に対してあらかじめ用意されたラベルを選択して付与する問題の一種である。具体的には日英の音声対話書き起こしデータに対して発話意図のタグを付与する問題を対象とする。そして統計的な枠組みに基づいた手法を提案する。

すでに音声翻訳システムの音声認識、あるいは翻訳の性能を向上させる目的で、確率的に発話意図を予測する研究が行われている [4, 5]。これら初期の研究では発話は適切な単位に分割されたものとして、タグを確率的に予測する手法を議論している。ところで、音声翻訳を対象とした場合、入力発話が適切に分割されている保証はない。そこで発話のある入力単位を設定した上でこれを最適に分割しタグを付与することが必要となる。

本稿では一話者の継続した発声、すなわちターンを入力とする。これは殆んどの音声対話システムで確実に検出できる単位である。そしてタグ付与の問題を日本語確率形態素解析と類似したモデルを利用して定式化する。この確率モデルは「言語情報」「音響情報」「状況情報」を一般的に包含しており拡張性が高い。本稿では日本語と英語のデータに対する実験を報告して提案手法の効果と問題点を議論する。

2 問題の設定

入力は一人の話者が継続して発声したターンとする。これに対する音声認識結果は正しく、形態素解析情報と話者情報が分かるものとする。ただし、句読点などの記号はないものとする。

このような情報を含んだ入力は一般に次のような 4 つ組の情報の系列だとみなすことができる。

$$\cdots(w_{i-1}, \boldsymbol{l}_{i-1}, \boldsymbol{a}_{i-1}, \boldsymbol{s}_{i-1}), (w_i, \boldsymbol{l}_i, \boldsymbol{a}_i, \boldsymbol{s}_i)\cdots$$

ここで w_i は表層形態素を表し、それぞれのベクトルは w_i に対する次のような情報を表す。

- $\boldsymbol{l}_i$: 言語特徴ベクトル。本稿では w_i の原型と品詞を格納する
- $\boldsymbol{a}_i$: 音響特徴ベクトル。本稿では w_i の後のポーズ時間（ms）を格納する
- $\boldsymbol{s}_i$: 状況特徴ベクトル。本稿では w_i の発声者を格納する

これにより、「お客」の *Hello I am John Phillips and ...* という発話に対しては

(Hello, (hello, INTER), 100, customer), (I,(i, PRON),0, customer)), (am, (be, BE), 0, customer) ....

という入力系列が得られることになる[1]。

以上の入力に対して本稿の課題は（1）入力系列を最適な数の発話意図（SA：speech act type）単位に分割すること、および（2）各 SA 単位に最適な SA タグを付与することである。また適切な数の SA タグ集合があらかじめ設定されているものとする。

本稿では SA 単位を u、その系列を $\boldsymbol{U}$ と記述する。また SA タグを t、その系列を $\boldsymbol{T}$ と記述する。また一般に x_s^e は記号 x の s から e までの系列を表すものとする。以上の記号を用いると問題は次のようになる：入力形態素列 $\boldsymbol{W}$ が与えられたときに最適な SA 単位系列 $\boldsymbol{U}$ およびその SA タグ系列 $\boldsymbol{T}$ を求めよ。この問題は単語グラフ上での確率的な探索の問題として捉えることができる。この単語グラフとは、入力系列が任意の形態素間で分割可能であるとして、その分割可能性すべてを保持したグラフである。以下、確率モデルと探索手法について説明する。

3 確率モデル

2 章の問題を確率モデルを使って定式化したものが式（1）である。すなわち入力系列 $\boldsymbol{W}$ が与えられた場合、その単語グラフ上で確率 $P(\boldsymbol{U}, \boldsymbol{T} \mid \boldsymbol{W})$ を最大にするような $(\hat{\boldsymbol{U}}, \hat{\boldsymbol{T}})$ を求める問題である。式（1）は式（3）に示すような 2 つの項に分解できる。第 1 項は、過去の SA 単位とタグの系列 $\boldsymbol{h}_j$（履歴 $\boldsymbol{h}_j$ =

[1] 以降では入力系列を単純に形態素列 $W = w_1, w_2, \ldots w_i, \ldots, w_n$ と記述するが、W は 4 つ組の系列であることに注意されたい。

u_1^{j-1}, t_1^{j-1}) が与えられたときに、以降の任意の形態素系列が一つの SA 単位 u_j を構成する確率である。第2項は同じ履歴が与えられたときに u_j が タグ t_j を持つ確率である。

$$(\hat{\boldsymbol{U}}, \hat{\boldsymbol{T}}) = \underset{\boldsymbol{U},\boldsymbol{T}}{\operatorname{argmax}} P(\boldsymbol{U}, \boldsymbol{T} \mid \boldsymbol{W}) \quad (1)$$

$$= \underset{\boldsymbol{U},\boldsymbol{T}}{\operatorname{argmax}} \prod_{j=1}^{k} P(u_j, t_j \mid \boldsymbol{h}_j, \boldsymbol{W}) \quad (2)$$

$$= \underset{\boldsymbol{U},\boldsymbol{T}}{\operatorname{argmax}} \prod_{j=1}^{k} P(u_j \mid \boldsymbol{h}_j, \boldsymbol{W}) \times P(t_j \mid u_j, \boldsymbol{h}_j, \boldsymbol{W}) \quad (3)$$

以後、第1項を「SA 単位の存在確率 P_E」と呼び第2項を「タグ付与確率 P_T」と呼ぶ。それぞれの確率をタグ付きのコーパスから有効に学習するため以下に述べる近似を行った。

3.1 SA 単位の存在確率の計算法

図 1 は 4 章で述べる探索中で、入力系列 w_1^{s-1} までの分割とタグ付与が終了しており w_s^{s+p-1} の部分系列に対して P_E と P_T を計算する状態を示したものである。式 (3) との対応に注意されたい。この図から P_E は入力と履歴が与えられた時に部分系列 w_s^{s+p-1} が一つの SA 単位になる確率であることがわかる。これを有効に推定するため次のように近似した。

$$P_E \simeq P(B_{w_{s-1},w_s} = 1 \mid \boldsymbol{h}_j, \boldsymbol{W}) \quad (4)$$

$$\times P(B_{w_{s+p-1},w_{s+p}} = 1 \mid \boldsymbol{h}_j, \boldsymbol{W}) \quad (5)$$

$$\times \prod_{m=s}^{s+p-2} P(B_{w_m,w_{m+1}} = 0 \mid \boldsymbol{h}_j, \boldsymbol{W}) \quad (6)$$

ここで $B_{w_x,w_{x+1}}$ は形態素 w_x と w_{x+1} の間に SA 単位の境界が存在する場合に値 1、しない場合に値 0 を取る確率変数である。P_E は系列 w_s^{s+p-1} の両端に境界が存在する確率と内部に境界がない確率の積で近似されている。この結果 P_E の推定は $P(B_{w_x,w_{x+1}} \mid \boldsymbol{h}_j, \boldsymbol{W})$ の推定に帰着する。

この推定には確率決定木を利用した。確率決定木はデータの希薄性に強く、離散属性や数値属性を容易に包含できる特徴がある。特に後者は入力系列として言語的特徴と音響的特徴を利用する著者らのモデルにとって有利である。また予備実験から条件部として $\boldsymbol{W}$ の一部 $\boldsymbol{W}' = w_{x-r+1}^{x+r}$ だけを使うことにした。これは着目している形態素境界周辺の $\pm r$ 個の入力系列情報である。以上を形式的に記述すると $P(B_{w_x,w_{x+1}} \mid \boldsymbol{h}_j, \boldsymbol{W}) \simeq P(B_{w_x,w_{x+1}} \mid \Phi_E(\boldsymbol{W}'))$ となる。右辺の記号 $\Phi_E(\boldsymbol{W}')$ は $\boldsymbol{W}'$ を同値なクラスに分類する決定木であり右辺全体で確率決定木を表現している。この確率決定木は $\boldsymbol{W}'$ を属性として、w_x と w_{x+1} の間での SA 境界の有無をクラスとした入力表を使って学習できる。決定木は C4.5 と同様、多重分岐で数値属性と離散属性を扱えるものを利用した。確率値はリーフのクラス分布で直接計算した。

3.2 タグ付与確率の計算法

P_T は以下の式で示すように確率決定木で近似した。

$$P_T \simeq P(t_j \mid \Phi_T(f(u_j), g(u_j), t_{j-1}, \ldots, t_{j-m})) \quad (7)$$

決定木 Φ_T には二つの関数 $f(u_j)$ と $g(u_j)$ と 過去 m 個のタグが使われている。$f(u_j)$ は形態素列 u_j の話者を返す関数である。一方 $g(u_j)$ は形態素列から次に説明する手がかり語群を抽出する関数である。

日本語の「ですか」は発話意図「疑問」を表しやすいなど、一般に SA タグは形態素との相関が高い。そこで発話中の形態素を使って P_T を推定すると有効である。しかしすべての形態素が発話意図と相関が高いわけではない。そこであらかじめ各タグと相関の高い形態素 10 語を正解コーパスから χ^2 値を使って抽出して手がかり語リストを作成しておく。$g(u_j)$ は u_j の形態素列中から手がかり語リストに合致するものを抽出する[2]。決定木を学習する際には手がかり語リストそのものを属性として、出現した形態素の属性値を 1 とする。

4 探索

探索には確率日本語形態素解析のために開発されたアルゴリズムを利用した [3]。日本語形態素解析はべた書き文字列を最適に分割して品詞を付与する作業である。これに対する本稿の問題は、文字列の代りに形態素列を入力として、品詞の代りに SA タグを付与する問題と見ることができる。すなわち両者は形式的に同型の問題である[3]。このためこのアルゴリズムをほぼそのまま利用できる。このアルゴリズムは任意の個数の過去のタグを確率モデルに利用できるため式 (7) との整合性が良い。また後ろ向きに A* 探索を行うことで n 最良解探索をできる利点もある。

5 実験

5.1 実験データ概要

実験に使ったタグ付データについて説明する。元になったデータは旅行手配に関する日本語話者と英語話者

[2] 比較は原型で行う。

[3] 図 1 の w が文字で u_j は辞書引きで得られる形態素候補と考えるとよい。ただし本稿の問題は辞書がない。これが主要な相違である。

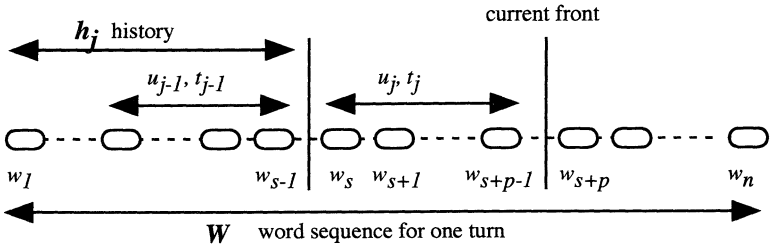

図 1: 確率計算法

表 1: データの概要

Counts	Japanese	English
Turn	2,020	2,020
SA unit	5,416	4,675
Morpheme	38,418	27,639
SA tag type	30	19

の通訳を介した会話である。このため会話の書き起こしには同じ内容の日本語と英語がある。また書き起こしには正しい形態素情報が付与されている [2]。

これらの会話データの 95 ファイルに対して発話の分割とタグ付与を人手で行った。日本語タグセットは [6] で提案されたものであり英語のタグセットはこれを縮小したものである。表 1 に利用したタグ付データの内容を示す。 日本語の 1 ターンの平均分割数は 2.68 で英語は 2.31 である。また日本語の 1 ターンの平均形態素数は 18 で 英語は 12.7 である。タグの種類は日本語が 30 で英語が 19 である。これらの数字からは英語のタグ付与の方が簡単だと予想される。

5.2 評価法

タグ付与結果はラベル付きの括弧の系列と見ることができる。そこで評価には「ラベルつき括弧付けの照合」を利用した [3]。概要は以下の通りである。正解データの出力括弧数を R とする。システムの出力括弧数を S とする。また照合した括弧数を M とする。以上の数から再現率 M/R と 適合率 M/S を計算できる。ここで次の 2 種類の照合を考える。一つは「分割照合」であり、開き括弧と閉じ括弧の位置が一致したものを照合と見なす。もう一つは「分割タグ付与照合」であり、分割照合の条件を満たした上でタグが一致したものを照合と見なす。分割照合は、開始と終了位置が同時に正しいものを正解としており、片方だけの評価よりも厳しくなっている。分割タグ付与照合は分割照合の性能に押さえられることに注意されたい。

5.3 結果

ここでは紙数の関係から日本語のデータに対して行った実験を中心に述べる。タグ付与に使う確率は P_E と P_T の二つの項からなる。この確率は 3.1 節と 3.2 節に示した決定木 Φ_E と Φ_T を使って推定しているが、学習パラメタには変更の余地がある。そこでこれらの効果を次の 2 種類の実験で確認した。

(1) SA 単位存在確率 P_E 推定へのパラメタの影響
P_T のパラメタを固定して P_E のパラメタを変更する。結果は「分割照合」で評価する。

(2) タグ付与確率 P_T 推定へのパラメタの影響
P_E のパラメタを固定して P_T のパラメタを変更する。結果は「分割タグ付与照合」で評価する。

日本語に対する種類 (1) の実験について述べる。この時の P_T 推定用の確率決定木のパラメタは $f(u_j)$、$g(u_j)$、t_{j-1} すなわち、「話者」「手がかり語」「一つ前の SA タグ」に固定した。この設定下で下記パラメタ群を属性とした P_E の確率決定木を学習した。

(A): 形態素 w_x^{x+1} の表層形と品詞。すなわち $r = 1$
(B): 形態素 w_{x-1}^{x+2} の表層形と品詞。すなわち $r = 2$
(C): (A) に w_x と w_{x+1} の間のポーズを追加
(D): (B) に w_x と w_{x+1} の間のポーズを追加

表 1 の日本語データを対象に (A–D) の各場合についてタグ付与実験を行った。以下の実験結果はすべて 10 回の交差確認法によって得たものである。「分割照合」に関する結果を表 2 に示す。この表に示すように再現率、適合率ともに (A) から (B) (C) (D) へ上昇した。またこれらの平均値の有意差を t 検定で評価した[4]。再現率については (A) と (B) (C) (D) の各間に有意差を検出できた。しかしその他の組み合わせに有意差はなかった。一方、適合率はついてはどの組み合わせにも有意差はなかった。一般に再現率が上昇すると

[4] 10 回の試行で得られた平均値を 2 標本検定の考え方で検定した。自由度 18 で両側有意水準 5% の t 値は 2.01 である。ここではこれより大きな t 値の場合に有意差があると認定した。

表 2: 分割照合結果

Parameter	Recall rate %	Precision rate %
(A)	89.50	91.99
(B)	91.89	92.92
(C)	92.00	92.57
(D)	92.20	92.58

表 3: 分割タグ付与照合結果

Parameter	Recall rate %	Precision rate %
(E)	72.25	72.70
(F)	74.91	75.35
(G)	74.83	75.29
(H)	74.50	74.96

適合率は下がる場合が多い。これに対して有意差はないものの適合率が上昇したことを考えると (A) から (B) (C) への条件の変更は有効だったと結論できる。すなわち形態素境界前後の 1 形態素の表層と品詞だけを使うより、この範囲を 2 に広げる、あるいはポーズを利用すると SA 単位の認定に有効である。ただし両者を使っても効果はない。

次に種類 (2) の実験結果を述べる。ここでは P_E 推定用確率決定木の学習パラメタを条件 (C) に固定した。この設定で下記のパラメタを使った P_T の確率決定木を学習した。

(E): u_j 中の手がかり語。すなわち $g(u_j)$

(F): (E) に t_{j-1} を追加

(G): (E) に t_{j-1} と t_{j-2} を追加

(H): (E) に t_{j-1} と話者情報 $f(u_j)$ を追加

これらの実験の「分割タグ付与照合」の結果を表 3 に示す。再現率、適合率ともに (E) から (F) (G) (H) へ上昇している。これらの平均値の検定を行った結果、再現率、適合率とも (E) と (F)、(E) と (G) の間に有意差を検出できた[5]。一方、その他の組み合わせでは有意差を検出できなかった。以上の結果から、タグの付与に関しては一つ前のタグ t_{j-1} を追加することは有効だが、さらに前のタグ t_{j-2} や話者情報を加えても性能は改善されないと結論できる。

以上とほぼ同様の実験を英語に対して実施した[6]。種類 (1) の実験では P_T を (H) に固定した。この実験では条件 (B) が最良であり「分割照合」の再現率 71.92%、適合率 78.10% を得た。種類 (2) の実験では P_E を (B) に固定して日本語と同様の実験を行った。この結果 (H) で最良の結果、「分割タグ付与照合」の再現率 53.1% と適合率 57.75% を得た。以上の結果は 5.1 節の予想に反して日本語よりかなり低い。特に分割照合の性能が低いために分割タグ付与照合の性能が悪化している。追加実験を行ったところ英語ターンの分割に品詞情報が日本語ほど効かないことがわかった。これが日英の差の主な原因であった[7]。この対策は今後の課題である。

[5] こちらは両側有意水準 10% である。すなわち $t > 1.73$。

[6] ポーズを使った実験は行わなかった。

6 おわりに

確率モデルに基づいた発話の分割と意図認識の手法を提案し、日本語と英語のデータを使った実験を報告した。提案手法の確率モデルは今回使った情報以外を利用できるようになっている。今後は [1] のように音響的な情報を多用する研究も行いたい。

参考文献

[1] M. Cettolo and D. Falavigna. Automatic detection of semantic boundaries based on acoustic and lexical knowledge. In *Proceedings of ICSLP-98*, Vol. 4, pp. 1551–1554, 1998.

[2] T. Morimoto, N. Uratani, T. Takezawa, O. Furuse, Y. Sobashima, H. Iida, A. Nakamura, Y. Sagisaka, N. Higuchi, and Y. Yamazaki. A speech and language database for speech translation research. In *Proceedings of ICSLP '94*, pp. 1791–1794, 1994.

[3] M. Nagata. A stochastic japanese morphological analyzer using a forward-DP and backward-A* N-best search algorithm. In *Proceedings of Coling94*, pp. 201–207, 1994.

[4] M. Nagata and T Morimoto. An information-theoretic model of discourse for next utterance type prediction. *Transactions of Information Processing Society of Japan*, Vol. 35, No. 6, pp. 1050–1061, 1994.

[5] N. Reithinger and E. Maier. Utilizing statistical dialogue act processing in verbmobil. In *Proceedings of the 33rd Annual Meeting of the ACL*, pp. 116–121, 1995.

[6] M. Seligman, L. Fais, and M. Tomokiyo. A bilingual set of communicative act labels for spontaneous dialogues. Technical Report TR-IT-0081, ATR-ITL, 1994.

[7] 英語は分割点前後の品詞に制限が少ない。