表層情報を用いた統計的手法による発話文の自動分類

青柳達也　　山本幹雄　　板橋秀一

筑波大学
〒305　茨城県つくば市天王台1-1-1

あらまし
現在、音声コーパスに各種のタグを付与し、音声認識器の性能向上や対話のモデル化などに利用する試みが行われている。しかし、タグを付与する際に、付与者によって揺らぎが生じたり、また、大量のデータにタグを付与する作業が大変な作業であるなど、問題点が多い。そこで、本研究では、付与者の支援となるようなタグの自動推定手法の確立を目的とした予備実験について述べる。具体的には、発話文の表層情報（話者、文末形式、文末での韻律）、及び発話文タイプのbigramを用いて、統計的手法により発話文タイプを推定する手法を提案する。結果としては、上位1つのみの場合、開いた実験で76.9%、閉じた実験で93.0%の正答率を、上位3個までなら開いた実験で89.4%、閉じた実験で99.8%の正答率を得た。
キーワード：発話文、表層情報、自動分類、発話文タイプ

A statistical classification method of utterances using surface information

Tatsuya Aoyagi　　Mikio Yamamoto　　Shuichi Itahashi

University of Tsukuba
1-1-1　Tennoudai, Tsukuba-shi, Ibaraki 305 JAPAN

Abstract
Illocutionary force type(IFT) information is useful for modeling dialogue structure and improving speech recognition system. However it is hard to build large corpora annotated with IFT by hand. In this paper, we describe a preliminary experiment to develop a semi-automatic IFT annotation system that facilitates work of human annotators. The method determines IFTs using end forms and prosody of input utterances, and previous IFTs that are important information sources about speaker's intention. We used bigram for modeling sequences of end forms and IFTs, and quantification theory type II for modeling prosody. It achieved 76.9% accuracy using test data and 93.0% using training data for the top candidate.
keywords:utterance, surface information, automatic annotating, illocutionary force type

1　はじめに

音声コーパスに各種のタグを付与し、それを音声認識での候補の絞り込や、認識誤りの回復、さらには対話のモデル化に用いる研究が行われている[1, 2, 3]。タグには韻律や形態素などの表層的なものから、文の機能、話者の意図などを表わすものなど、様々な階層がある。ここで、階層が深くなればなるほど、付与者の主観がタグの決定に影響し、揺らぎが生じる。そこで、本研究では、そのようなタグをある程度まで自動的に付与・選択し、タグ付けの支援とるような手法を提案する。

発話文の機能を機械で推定するためには、韻律や文形式などの表層情報と、意味や話題などの文脈情報・深層情報が必要である。これらを機械で処理する場合、前者は比較的扱いやすいが、後者を扱うのは困難である。そのために、本研究では、発話文の機能を表わすタグ（発話文タイプと呼ぶ）、話者、及び文末形式の組み合わせの遷移確率モデルを、韻律情報と組み合わせることにより推定を行った。

2 対話データ

対話データは、文部省重点領域研究「音声対話」の対話音声コーパスVol.1中の8対話を用いた。諸元を表1に示す。これらは主に2つのトピック、案内タスクとテレフォンショッピングタスクに別れており、話者はそれぞれ5人と4人の合計で9人となっている。基本的には、一方の話者が質問し、もう一方がそれに答える、「質問－応答」対話である。表1の発話単位は、ポーズで挟まれた句音声（一区切りとして発生された音声区間）となっている。これは、音声認識器の学習用に人手で付与された発話区切りのタグをそのまま用いたためである。なお、実際の実験では確率遷移モデルの構築には全発話単位（665個）を用い、発話文タイプの推定実験ではF_0のデータが抽出できなかった発話単位は除外して実験を行った。

表1　対話データ

対話番号	トピック	発話単位数
osa0018	予約案内（スキー旅行）	155(125)
osa0019	地理案内（厚生年金会館）	71(67)
osa0020	地理案内（厚生年金会館）	43(33)
osa0021	地理案内（厚生年金会館）	36(18)
tsu1103	テレフォンショッピング	94(88)
tsu1107	テレフォンショッピング	90(90)
tsu1204	テレフォンショッピング	77(76)
tsu1208	テレフォンショッピング	99(94)
合計		665(591)

※括弧内は評価実験に用いた発話単位数

3 発話文タイプの推定に用いる情報

ここでは発話文タイプの推定に用いた各情報について述べる。発話の表層情報として話者、文末形式、及び韻律情報を、文脈情報として、直前の発話単位の発話文タイプを用いた。

3.1 韻律情報

発話文の機能による韻律の違いは、その文末の部分に良く現われる[4]。そこで、文末形式を持つ発話単位に関しては、その文末形式の部分に対して、持たない場合は発話単位全体に対して、話者ごとに正規化したF_0及びパワーに対して粗い折線近似を行い、折線の本数及び最後の1本の折線の傾きと開始点・終了点でのF_0及びパワーの値の合計8個の情報を特徴量として用いた。折線近似は、以下のアルゴリズムで行った。

STEP1　近似したいパターンの始点と終点を直線で結ぶ
STEP2　直線（折線）とパターンの各点間の距離を計算する。
STEP3　距離が閾値以下であったなら終了。閾値以上であったなら、もとの直線の始点・終点と最大距離の点を結び、折線を分割する。
STEP4　すべての直線に対して、STEP2～3を繰り返す。

閾値の設定は、実験的に視察により行った。折線近似の例として、F_0の折線近似の例を図1に示す。粗い近似を行うことにより、F_0抽出時におけるノイズの影響を除去することができ、おおまかな韻律の変化の様子を捉えることが可能である。折線の傾きはF_0では昇調・降調に、パワーでは声の大きさの変化に関係する。F_0及びパワーの開始点・終了点での値が疑問調の知覚に関係するとの報告が有るので[4]、パワーでもその可能性が十分あると考え採用している。また、これらの値は傾きと組み合わせることにより、発話時間を表わすことができる。折線の本数は、その文末形式（文末形式がない場合は発話単位全体）の長さを大まかに反映するので、これにより、文末が伸びるような発話の特徴を捉えることができる。

3.2 文末形式

実際に対話データ中に出現した文末形式は、図2の計23種類であった。それらの中で、「で」、「です」、「ですか」、「ます」、「文末形式なし」の5つが、2つ以上の発話文タイプを持ち、文末形式だけでは発話文タイプ

を決定できない、曖昧なものであった。それらがどのような曖昧性を持つかを表2に示す。括弧の中は、発話単位数である。可能性のある発話文タイプが最も多いのは「文末形式なし」で９個、最も少ないのが「ですか」で３個、平均では5.4個であった。全発話単位では、文末形式のみで発話文タイプが一意に決まるものが全体の48%、曖昧なものが52%であった。これより、半数近くは文末形式以外の情報を用いる必要があることがわかる。

3.3 発話文タイプ

文の機能を表わすタグとして、全部で１０種類の発話文タイプを定義し、対話データに付与した（表3）。タグは階層構造になっており、「／」で階層の区切りを表わしている。今回はなるべく揺らぎが生じないように、「質問」と「説明」の２つのタグにしか階層を定義せず、さらに２階層までしか用いていない。それにもかかわらず、２階層目では判断に困ることが多々あった。特に「説明」では、「提示」と「意思表示」の両機能を持つようなものが多く見られた。そのような場合には、前後の文脈を

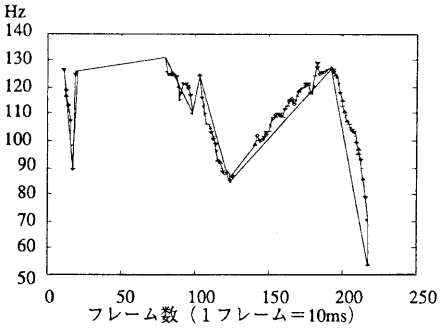

図１　F_0の折線近似の例

で　でした　でしたっけ　でしたら
でしょうか　です　ですが　ですか
ですから　ですけど　ですけれども
ですね　ですよ　ください　ました
ましたっけ　ます　ますが　ますか
ますけど　ますけれども　はい
文末形式なし

図２　出現した文末形式

表２　発話文タイプに対して曖昧性のある文末形式

文末形式	個数	発話文タイプ
で	31	説明(2)　説明／意思表示(1)　説明／提示(4)　質問/yes-no(24)
です	41	説明(8)　説明／提示(16)　挨拶(4)　相槌応答(13)
ですか	29	質問/yes-no(22)　質問／説明要求(2)　相槌応答(5)
ます	82	説明(21)　説明／意思表示(10)　説明／提示(13)　質問／説明要求(23) 挨拶(11)　相槌応答(4)
文末形式なし	121	説明(19)　説明／意思表示(1)　説明／提示(28)　質問／yes-no(5) 質問／説明要求(25)　復唱(25)　挨拶(7)　相槌応答(11)

※括弧内は発話単位数を表わす

表３　発話文タイプの定義

発話文タイプ	定義	個数	例
質問／yes-no	yes/no質問	89	「何か宿空いているでしょうか」
質問／説明要求	相手に説明を要求する質問	84	「商品名をお願いします」
説明／提示	新しく何かを提示する説明	73	「商品名はウッドラック」
説明／意思表示	自分の意思を表明する説明	28	「空港に行きたいんですけども」
説明	上記以外の説明	67	「駅の一番北側の出口を出ます」
復唱	相手の発話を繰り返す場合	25	「ウッドラック」
挨拶	挨拶	29	「失礼します」
相槌応答	相槌に近い応答	44	「ああ、そうですか」
命令	命令	4	「右に曲がって下さい」
はい	はい	152	「はい」

考慮し、話者の比重が重く置かれていると判断された方を採用した。それでも決定できない場合は、2階層目のない「説明」としてある。

4 発話文タイプの推定手法

ここでは用いた推定手法について述べる。基本的にはシンボルデータ（話者／文末形式／発話文タイプ）の組み合わせの遷移確率モデルを用い、韻律情報に関しては一般化された数量化理論第II類[5]を用いた。タイプを推定すべき発話単位の文末形式がデータとして与えられており、かつそれだけで発話文タイプが一意に決定できる場合は、文末形式のみで決定した。

4.1 遷移確率モデル

学習データより、話者、文末形式、及び発話文タイプの組み合わせに対する条件付確率を求めた。各式で用いている変数は、以下の通りである。なお、uは現在の発話単位の情報のみを用いるという意味でunigram、bは直前の発話の情報も加えて使用するという意味でbigramと呼ぶことにする。

t_i ：現在の発話単位の発話文タイプ
s_i ：現在の発話単位の話者
l_i ：現在の発話単位の文末形式
t_{i-1}：直前の発話単位の発話文タイプ
s_{i-1}：直前の発話単位の話者
l_{i-1}：直前の発話単位の文末形式

(1) 発話文タイプからの推定

$$u = p(t_i) \tag{1}$$

$$b = p(t_i \mid t_{i-1}) \tag{2}$$

(2) 文末形式と発話文タイプからの推定

$$u = p(t_i \mid l_i) \tag{3}$$

$$b = p(t_i \mid t_{i-1}, l_{i-1}, l_i) \tag{4}$$

(3) 話者と発話文タイプからの推定

$$u = p(t_i \mid s_i) \tag{5}$$

$$b = p(t_i \mid t_{i-1}, s_{i-1}, s_i) \tag{6}$$

(4) 話者、文末形式、及び発話文タイプからの推定

$$u = p(t_i \mid s_i, l_i) \tag{7}$$

$$b = p(t_i \mid t_{i-1}, s_{i-1}, l_{i-1}, s_i, l_i) \tag{8}$$

4.2 数量化理論第II類による韻律情報からの発話文タイプの判別

一般化された数量化理論第II類[5]により、韻律情報を用いて発話文タイプの推定を行う。数量化理論第II類の外的基準として、3つの情報（現在の発話単位の話者／文末形式／発話文タイプ）を組み合わせたものと、発話文タイプのみの2種類を用いた。内的基準としては、3.1節で述べた8つの韻律情報を用いた。

発話文タイプを推定する方法としては、前述のように3通りの方法を用いた。ここでは、数量化理論第II類のみの方法と、遷移確率モデルと数量化理論第II類を組み合わせる方法について説明する。まず、数量化理論第II類のみの場合には、判別に必要なパラメータは学習データから獲得し、このパラメータを用いてテストデータの外的基準を推定した。遷移確率モデルと数量化理論第II類の結果を組み合わせる場合には、数量化理論第II類の学習データにおける判別率を利用した。例えば、学習データを数量化理論第II類の結果を利用して再判別を行ったときに、Xと判定されたデータのうちY%が実際にXであったならば、その外的基準についての判別率はY%であるとした。

学習に関しては、モデル（3）とモデル（4）では、文末形式を情報として用いているので、学習は各文末形式ごとに行った。ただし、「で」「です」「ですか」については、学習データが少ないので、「で」と「です」、「です」と「ですか」の2つのセットにまとめて判別率を求めた。この場合、「です」に対する判別は、「で」と「です」をまとめたセットで行った。モデル（1）とモデル（2）に関しては、すべての発話単位で判別率を求めた。

4.3 各手法の結合

(9)式で与えられるPを最大にする発話文タイプを、その発話単位の発話文タイプとして決定する。

$$P = w_1 u + w_2 b + w_3 q \quad (w_1 + w_2 + w_3 = 1) \quad (9)$$

重みに関しては、制約式を満たすよう各重みを0.1刻みで変化させ、学習データに対して最も正答率が高かったものを採用した。なお、uはunigram、bはbigram、qは数量化理論第II類による判別率を表わす。

5 実験

5.1 実験手法

F_0を抽出できた全発話（591発話単位）より、100個の発話単位をテストデータとしてランダムサンプリングし、残りを学習データとした。各データセット中の、文末形式から一意に発話文タイプを決定できる発話（曖昧性なし）の個数は表4の様になった。

5.2 結果

実験は各情報と各手法の使用、不使用の組み合わせを変えて行った。結果を表5に示す。unigramに関しては、正答率の向上に寄与しなかったので、表には載せていない。なお、正答率は上位3個まで求めた。モデル（4）、モデル（3）に関しては、文末形式を情報として用いていており、それによって発話文タイプが一意に決定できるものが含まれているので、曖昧性があるものだけについての正答率（1位のみ）も挙げてある。それらに対し、文末形式を情報として用いないモデル（2）とモデル（1）に関しては、すべての発話単位が曖昧なものとなっているので、「曖昧のみ」欄は空欄となっている。

表4　学習データとテストデータ

	学習データ	テストデータ
曖昧性なし	255	52
曖昧性あり	236	48
合計	491	100

※単位は発話単位

表5　発話文タイプの推定結果

モデル	情報	手法	1位	2位	3位	曖昧のみ	w2	w3
（4）	話者　文末形式 発話文タイプ	B Q	76.0/93.0	82.7/97.8	89.4/99.8	51.9/87.5	0.9	0.1
		B -	74.0/89.9	82.7/96.9	86.6/98.9	48.1/82.0	1.0	-
		- Q	70.2/82.6	77.9/94.7	82.7/98.9	40.4/69.2	-	1.0
（3）	文末形式 発話文タイプ	B Q	76.0/89.7	85.6/98.0	91.4/98.5	51.9/81.6	0.9	0.1
		B -	71.2/87.0	84.6/95.4	89.4/97.6	42.3/76.9	1.0	-
		- Q	67.3/75.8	74.0/88.4	81.7/95.6	34.6/56.9	-	1.0
（2）	話者 発話文タイプ	B Q	28.9/44.7	50.0/68.6	65.4/83.1	――	1.0	0.0
		B -	28.9/44.7	50.0/68.6	65.4/83.1	――	1.0	-
		- Q	23.1/25.5	36.5/30.2	42.3/50.2	――	-	1.0
（1）	発話文タイプ	B Q	28.9/39.6	42.3/61.2	61.5/78.4	――	0.6	0.4
		B -	21.2/37.7	46.2/56.1	63.5/70.1	――	1.0	-
		- Q	23.1/23.5	36.5/47.5	59.6/65.5	――	-	1.0

※手法欄のBはbigram、Qは数量化理論第II類を用いることを表わす。
※1位、2位、3位は、それぞれ上位1位、2位、3位までの結果を用いた正答率である。
※w2、w3、はそれぞれbigram、数量化理論第II類の結果への重みである。
※モデル（4）とモデル（3）では、文末形式のみで発話文タイプが一意に決定できるものが含まれているので、一意に決定できない、曖昧なものに対しての正答率（1位のみ）も示してある。モデル（2）とモデル（1）には、そのような曖昧性のないものは含まれていないので、「曖昧性のみ」の項目は空欄になっている。
※数量化理論第II類を用いるときは、「情報」の項目にはとして「韻律情報」が追加される。

5.3 考察

すべての情報と、bigram、数量化理論第II類を組み合わせた場合、上位1つのみの場合で、開いた実験で76.0%、閉じた実験では93.0%の正答率であった。また、上位3個まで考えるならば、開いた実験で89.4%、閉じた実験で99.8%の正答率を得た。

モデル（4）とモデル（3）、あるいはモデル（2）とモデル（1）を比較すると、話者情報を用いると正答率が向上することがわかる。これは、今回用いた対話データのような「質問ー応答」といった定型的な対話では、話者によって発話文タイプに偏りがある、つまり基本的に質問する側と応答する側に決まっていることが原因であると思われる。同様に、モデル（4）とモデル（2）、あるいはモデル（3）とモデル（1）を比較すると、文末形式を用いると飛躍的に正答率が向上することがわかる。これは、今回の実験では、文末形式のみで発話文タイプが一意に決まるものがデータの半数近くあったためであり、文末形式と発話文タイプの関連が大きいことがわかる。

手法に関しては、単独ではbigramが最も良い結果を示している。特にモデル（2）では、bigramのみを用いた場合の正答率が最も良くなっている。

数量化理論第II類（韻律情報）を用いる場合と用いない場合とでは、各モデルにおけるbigramだけを用いる場合とbigramと数量化理論第II類を組み合わせた場合の比較から、数量化理論第II類が正答率の向上に若干ながら寄与していることがわかる。そこで、文末形式「です」を持つ発話単位のみでの実験を行った結果を表6に示す。

表6　推定結果（文末形式「です」のみ）

情報	手法	1位
話者　文末形式	B　Q	72.7/96.7
発話文タイプ	B　-	63.6/96.7

表5の「曖昧のみ」の項目と表6を比較すると、「です」は曖昧性を持つ文末形式の中でも、正答率が高くなっていることがわかる。さらに、数量化理論第II類によって、テストデータの正答率が9%向上している。これは、文末形式によっては、韻律情報が発話文タイプを決定するのに有効な情報になることを示唆している。

6 まとめ

表層情報である話者、文末形式、韻律情報と、深層的な情報である発話文タイプを用いることで、タグ付けの支援となるような、発話文タイプの自動推定手法を提案した。すべての情報を用い、bigram、数量化理論第II類を組み合わせることで、上位3個で、開いた実験で89.4%、閉じた実験で99.8%の正答率を得ることができた。

今後の課題としては、さらにデータ数を増やす必要がある。また、韻律情報に関しては、どのような場面で韻律情報を用いると正答率の向上に効果があるのか、またはどのような韻律情報を用いれば良いのかを含め、さらなる検討を行う必要がある。

参考文献

[1] 石崎雅人, 小磯花絵,「対話研究の新しい流れ」, 人工知能学会研究資料, SIG-SLUD-9503-1, (1996)

[2] 北研二, 福井義和, 永田昌明, 森元逞,「発話タイプ付きコーパスを用いた確率的対話モデルの自動生成」, SIG-SLUD-9503-8, (1996)

[3] 永田昌明, 鈴木雅美,「日英対話コーパスへの発話行為タイプ付与の試みとその統計的対話モデルへの利用」, 人工知能学会研究資料, SIG-SLUD-9302-7, (1993)

[4] 板橋秀一, "文末のピッチコンターと疑問調の知覚について -日本人とスウェーデン人の応答の比較-", 音講論, 2-2-3, pp.549-550, (1979.6)

[5] 小林龍一,「数量化理論入門」, 日科技連出版, (1981)