

# 対話行為に固有の特徴を考慮した自由対話システムにおける 対話行為推定

福岡 知隆<sup>†,††</sup>・白井 清昭<sup>†</sup>

対話行為の自動推定は自由対話システムにおける重要な要素技術のひとつである。機械学習を用いた既存の対話行為の推定手法では、機械学習に用いる特徴のセットを1つ設定するが、この際に個々の対話行為の特質は十分に考慮されていなかった。機械学習の特徴の中にはある特定の対話行為の分類にしか有効に働かないものもあり、そのような特徴は他の対話行為の分類精度を低下させる要因になりうる。これに対し、本論文では対話行為毎に適切な特徴のセットを設定する。まず、28個の初期の特徴を提案する。次に、対話行為毎に初期特徴セットから有効でない特徴を削除することで最適な特徴セットを獲得する。これを基に、入力発話が対話行為に該当するかを判定する分類器を対話行為毎に学習する。最後に、個々の分類器の判定結果ならびに判定の信頼度から、適切な対話行為をひとつ選択する。評価実験の結果、提案手法は唯一の特徴セットを用いるベースラインと比べてF値が有意に向上したことを確認した。

キーワード：自由対話，対話システム，対話行為，機械学習，特徴選択

## Dialog Act Classification Using Features Intrinsic to Dialog Acts in an Open-Domain Conversation

TOMOTAKA FUKUOKA<sup>†,††</sup> and KIYOAKI SHIRAI<sup>†</sup>

The classification of dialog acts of user's utterance is one of the important fundamental techniques in open-domain conversational systems. Most previous studies on the classification of dialog acts were based on supervised machine learning; however, the characteristics of individual dialog acts were not considered. Some features for machine learning may increase the accuracy of classification for a particular dialog act, whereas decrease the accuracy for other dialog acts. In this study, an appropriate feature set is defined for each dialog act to improve the performance of the classification of the dialog acts. First, 28 features are proposed as an initial set. Second, for each dialog act, an optimal set of the features is identified by removing ineffective features from the initial set. Third, binary classifiers that judge whether a dialog act is suitable for a given utterance are trained using the optimized feature set. Finally, one dialog act is chosen based on the results provided by the binary classifiers. The reliability of the judgment of the binary classifiers is also considered. Results of

---

<sup>†</sup> 北陸先端科学技術大学院大学 先端科学技術研究科, School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology

<sup>††</sup> 現在, 株式会社 Nextremer, Presently with Nextremer Co., Ltd.

experiments showed that our proposed method significantly outperformed a baseline that was trained using a single feature set.

**Key Words:** *Open Domain Conversation, Dialog System, Dialog Act, Machine Learning, Feature Selection*

## 1 はじめに

近年、対話の内容を特定のタスクに限定しない自由対話システムの研究が盛んに行われている (Libin and Libin 2004; Higashinaka, Imamura, Meguro, Miyazaki, Kobayashi, Sugiyama, Hirano, Makino, and Matsuo 2014). 対話システムの重要な要素技術の1つにユーザの発話の対話行為の自動推定がある。対話行為の推定は自由対話システムにおいて重要な役割を果たす。例えば、対話行為が「質問」の発話に対しては知識ベースから質問の回答を探して答えたり、映画の感想を述べているような「詳述」の発話に対しては意見を述べたり単にあいづちを返すなど、対話システムは相手の発話の対話行為に応じて適切な応答を返す必要がある。

対話行為の推定手法として機械学習を用いた手法が既に提案されている (Milajevs and Purver 2014; 磯村, 鳥海, 石井 2009; 関野, 井上 2010; Kim, Cavedon, and Baldwin 2010; 目黒, 東中, 杉山, 南 2013). しかし、機械学習に用いる特徴<sup>1</sup>を設定する際、個々の対話行為の特質が十分に考慮されていないという問題点がある。既存研究の多くは、対話行為の自動推定を多値分類問題と捉え、対話行為の分類に有効と思われる特徴のセットを1つ設定する。しかし、機械学習の特徴の中には、ある特定の対話行為の分類にしか有効に働かないものもある。例えば、ユーザの発話の対話行為が(質問に対する)「応答」であるかを判定するためには、発話者が交替したかという特徴は重要だが、対話行為が「質問」であるかを判定するためには、相手の発話の後に質問することもあれば自身の発話に続けて質問することもあるので、話者交替は重要な特徴とは考え難い。

本論文では、上記の問題に対し、対話行為毎に適切な特徴のセットを設定することで個々の対話行為の推定精度を改善し、それによって全体の対話行為推定の正解率を向上させる手法を提案する。

## 2 関連研究

対話を形成する上で、話者の対話行為は対話の展開に強い影響を与えるだけでなく、対人印象や対人関係にも影響を及ぼしている (西田 1992). 自由対話においては、対話を継続するか否

---

<sup>1</sup> 本稿では、機械学習による識別のために用いる情報の種類(タイプ)のことを「特徴」、その具体的な情報のことを「特徴量」と呼ぶ。例えば、「単語 3-gram」は特徴、「思い+ます+か」はその特徴量である。

かの判断はユーザに委ねられており、対話の内容のみならず、対話システムの不自然な応答や不快な発話是对話の破綻に繋がる。自由対話を継続するには、ユーザと対話システムが良好な関係を築く必要があり、そのためには、話者の対話行為を正確に推測し、それに応じて適切な応答を返さなければならない。

## 2.1 対話行為の利用

対話システムにおける対話行為情報の利用目的として、ユーザ意図の理解、システムの応答文生成における条件、システムの対話制御などが挙げられる (Higashinaka et al. 2014; Inui, Ebe, Indurkha, and Kotani 2001; 前田, 南, 堂坂 2011; Sugiyama, Meguro, Higashinaka, and Minami 2013). 例えば、ユーザの発話を分析し、対話行為にクラス分けすることは、ユーザの意図理解の1つとみなせる。ユーザが挨拶をしているか、何かを質問しているのかなどをシステムが理解することで、その後の対話の展開を決定する。南らは行動予測確率に基づく報酬を設定する部分観測マルコフ決定過程 (POMDP) を用いた対話制御手法において、対話行為列の tri-gram による行動予測確率を導入した手法を提案し、その有効性を確認した (南, 東中, 堂坂, 目黒, 森, 前田 2012).

また、発話からの情報抽出のためのフィルタリング条件としても対話行為の情報は用いられる。平野らは、ユーザの発話からユーザ情報を抽出する手法を提案し、その手法では発話がユーザ情報を含むか否かを対話行為に基づき判断している (平野, 小林, 東中, 牧野, 松尾 2016).

## 2.2 対話行為推定

教師あり機械学習に基づく対話行為の自動推定では、機械学習に用いる基本的な特徴として単語 n-gram が利用されることが多い。これに加えて独自の特徴も提案されている。

単語 uni-gram は語順を考慮していないため、Milajevs らは単語 bi-gram を特徴として用い、単語 uni-gram のみよりも bi-gram を併用したときの方が高い精度が得られることを示した (Milajevs and Purver 2014). また、対話の流れを考慮するために前の発話の対話行為を特徴として利用し、その効果を評価した。磯村らは、頻度 2 以上の単語 uni-gram と単語 bi-gram、及び 1 つ前の発話の対話行為を特徴として、Conditinal Random Field (CRF) を用いて対話行為を推定し、75.77% の推定精度を得たと報告している (磯村 他 2009). 機械学習アルゴリズムとして Support Vector Machine (SVM) と Naive Bayes を用いた実験も行ったが、これらでは 1 つ前の発話の対話行為を特徴として利用しておらず、推定精度はそれぞれ 66.95% と 60.14% となり、CRF より劣る。関野らは、特徴として発話文字数、内容語数、発話順番を提案し、磯村の手法 (磯村 他 2009) の特徴にこれらを 1 つ以上追加したモデルを評価した (関野, 井上 2010). 全ての組み合わせにおいてその有効性が確認され、内容語数と発話順番を追加した場合が最も高い精度となった。Kim らは、bag-of-words に加え、対話中の話者の役割などの構造的な情報と、直前の発話

や同一話者によるこれまでの対話行為などといった対話の依存関係を機械学習の特徴として提案した (Kim et al. 2010). ドメインが限られた対話を評価の対象としているが, 96.86%という高い推定精度が得られている.

目黒らは, 多種多様な話題や語彙を含み, また非文法的な文が多いマイクロブログ中の発話に対する対話行為自動付与のため, シソーラスを用いて抽象化した単語 n-gram と文字 n-gram を特徴とする手法を提案した (目黒 他 2013). 評価実験の結果, Bag-of-Ngrams を特徴として用いたベースライン手法よりも高い精度を得た.

これらの先行研究では, 機械学習のために用いる特徴のセットは1つであり, それで全ての対話行為を推定している. しかし, どの特徴がどの対話行為の推定に有効に働くかなど, 特徴と対話行為の関係については議論されていない. 本研究では, 発話がある対話行為を持つか否かを推定する機械学習において, 対話行為それぞれに対して有効な特徴を自動的に選択する.

### 3 提案手法

本節では, 自由対話における発話を入力とし, その対話行為を推定する手法について述べる. 対話行為の分類クラスをあらかじめ定義し, その中から適切な対話行為のクラスを1つ選択する. 従来手法の多くは教師あり機械学習に基づくが, 学習のための特徴のセットはあらかじめ一律に定められている. しかし, 全ての特徴が全ての対話行為の分類に必要というわけではなく, ある特徴が特定の対話行為の分類に貢献しないことがある. そのような特徴は正解率を低下させる要因となりうる. この問題を解決するために, 提案手法では, 対話行為の分類クラス毎に異なる特徴のセットを設定する.

提案手法の処理の流れを図1に示す. 対話行為毎に, 入力発話がその対話行為に該当するかどうかを判定する二値分類器を学習する. その際, 対話行為毎に最適な特徴のセットを実験的に決める. また, 分類と同時に判定の信頼度も算出する. 次に, 二値分類器による判定の結果, ならびに判定の信頼度を基に, 入力発話の対話行為をひとつ選択する. 本論文では, 対話行為を選択するアルゴリズムとして, 3.5項で述べる4つの手法を提案する.

本論文では, 各対話行為の二値分類器をL2正則化ロジスティック回帰によって学習し, 学習ツールとしてLIBLINEAR (Fan, Chang, Hsieh, Wang, and Lin 2008) を用いた. LIBLINEARの学習パラメタはデフォルト値を用いた. 判定の信頼度はLIBLINEARが出力する確率を用いた.

#### 3.1 対話行為の定義

対話行為の定義としてはSWBD-DAMSL (Jurafsky, Shriberg, and Biasca 1997) が著名だが, かなり詳細な対話行為が定義されており, また自由対話を対象としたものではない. 自由対話を想定した対話行為のセット (Meguro, Minami, Higashinaka, and Dohsaka 2014) も提案されて

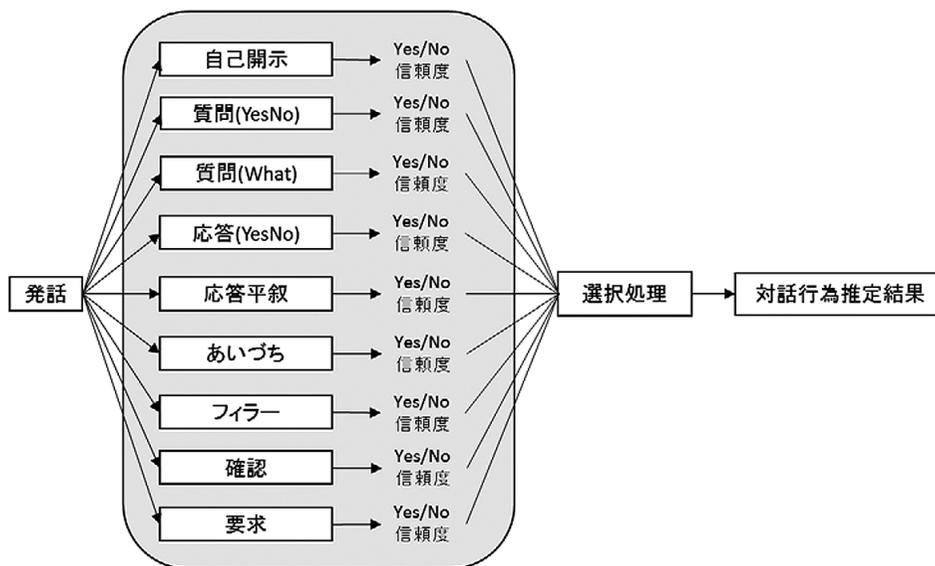


図 1 提案手法の流れ

表 1 対話行為の定義

$d_1$ 自己開示	発話者の考えや事実を述べる発話
$d_2$ 質問 (YesNo)	相手に対して「はい」「いいね」などの返答を求める質問
$d_3$ 質問 (What)	具体的な内容を問う質問
$d_4$ 応答 (YesNo)	「はい」「いいえ」に相当する短い応答
$d_5$ 応答 (平叙)	質問に対して具体的な内容を返す応答
$d_6$ あいづち	あいづちを表わす発話
$d_7$ フィラー	意味を持たないが間をつなぐための発話
$d_8$ 確認	相手が伝聞・理解したことを確認する発話
$d_9$ 要求	相手に対する何らかの要求を表わす発話

はいるが, 本研究では, 今後構築を目指す自由対話システムの仕様を考慮して独自に定義した 9 つの対話行為のセットを用いる. その一覧を表 1 に示す.

### 3.2 機械学習に用いる特徴

人間同士の自由対話を人手で分析し, 対話行為の言語的特徴を考慮して, 対話行為の推定に有効と思われる 28 個の特徴を設定した. その一覧を表 2 に示す. これらは大きく 4 つのグループに分けられる.

グループ 1  $f_1 \sim f_{10}$  は, 発話の内容を表わし, 全ての対話行為の分類に有効と考えられる特徴である. 現在の発話 (対話行為を推定すべき入力発話) ならびにその直前の発話に含まれる

表 2 対話行為推定の特徴

$f_1$ : 単語 n-gram	$f_{11}$ : 質問キーワード	$f_{21}$ : 話者交代
$f_2$ : 前発話の単語 n-gram	$f_{12}$ : 質問 (What) キーワード	$f_{22}$ : 自立語の有無
$f_3$ : 自立語	$f_{13}$ : 応答 (YesNo) キーワード	$f_{23}$ : 自立語の繰返しの有無
$f_4$ : 前発話の自立語	$f_{14}$ : あいづちキーワード	$f_{24}$ : 自立語繰返し 1
$f_5$ : 文末の単語 n-gram	$f_{15}$ : フィラーキーワード	$f_{25}$ : 自立語繰返し 2
$f_6$ : 前発話の文末単語 n-gram	$f_{16}$ : 文末要求表現	$f_{26}$ : 一単語発話 (自立語)
$f_7$ : 文末付属語列	$f_{17}$ : 文末あいづち表現	$f_{27}$ : 一単語発話 (非自立語)
$f_8$ : 前発話の文末付属語列	$f_{18}$ : 相手の過去の発話の対話行為	$f_{28}$ : 発話内単語繰返し
$f_9$ : 文末 n-gram ペア	$f_{19}$ : 話者の過去の発話の対話行為	
$f_{10}$ : 文末付属語列ペア	$f_{20}$ : 発話長	

単語 n-gram, 自立語, 文末に出現する単語 n-gram ならびに付属語の列を特徴とする.  $f_9, f_{10}$  は, それぞれ現在の発話と前発話の単語 n-gram, 付属語列の組を表わす. 単語 n-gram を用いた特徴 ( $f_1, f_2, f_5, f_6, f_9$ ) では  $n = 1, 2, 3$  とした.

**グループ 2**  $f_{11} \sim f_{17}$  は, 発話の内容を表わし, 特定の対話行為の推定に有効に働くと考えられる特徴である.  $f_{11} \sim f_{15}$  はそれぞれの対話行為の発話で頻出すると思われるキーワードである. これらのキーワードは訓練データを参照して人手で選定した.  $f_{16}$  は要求の発話の文末によく見られる表現であり, 文末が命令形の動詞, 動詞基本形+「な」の否定の命令形, 動詞連用形+「て」, 動詞連用形+「や」, これらの表現+「よ」or「ね」, のいずれかに当てはまることを表わす.  $f_{17}$  は, あいづちを示唆する文末表現「ね」が出現するかを表わす.

**グループ 3**  $f_{18} \sim f_{21}$  は, 発話の内容以外の情報を表わし, 全ての対話行為の分類に有効と考えられる特徴である.  $f_{18}, f_{19}$  は, それぞれ相手もしくは話者自身の直前のいくつかの発話の対話行為の列である. 対話行為列の長さは実験的に定める. 詳細は 3.3.2 で述べる.  $f_{20}$  は発話文中の文字数に基づく発話の長さである. 発話長を機械学習の特徴として用いる場合, 長さを適当な間隔 (1 ~ 5, 6 ~ 10, 11 以上, など) に切って発話長を分類するのが一般的であるが, その適切な間隔を決めるのは難しい. 本論文では, 「発話長が  $l \pm 2$  である」 ( $3 \leq l \leq 19$ ), 「発話長が 20 以上である」といった特徴量で発話長を表現する. 例えば, 発話長が 10 の発話に対しては,  $l = 8, 9, 10, 11, 12$  の特徴量の重みを 1 とする.  $f_{21}$  は現在と直前の発話の話者が同じかどうかを表わす. 実験に用いた自由対話コーパスでは, 同じ話者が 2 つ以上の発話を連続して発言することがあるため, この特徴を導入した.

**グループ 4**  $f_{22} \sim f_{28}$  は, 発話の内容以外の情報を表わし, 特定の対話行為の推定に有効に働くと考えられる特徴である.  $f_{22}$  の「自立語の有無」は自立語を含まなくても生成できる「応答 (YesNo)」, 「あいづち」, 「フィラー」とその他の対話行為の区別にも有効であると考えられる.  $f_{23} \sim f_{25}$  における「自立語繰返し」とは, 相手の前発話の自立語が現在の発話で繰返し用いら

れるかを表わす。単語を繰り返して聞き返す「確認」や、反復による「あいづち」の特徴を捉えられる。 $f_{23}$ は単純に自立語が繰り返されるか否かを考慮するが、より厳密に「確認」、「あいづち」を示唆する自立語の繰り返しを区別するため、繰り返される自立語が相手の前発話の文末に出現する場合を $f_{24}$ 、発話で繰り返される自立語が現在の発話における唯一の自立語である場合を $f_{25}$ とする。 $f_{26}$ 、 $f_{27}$ はそれぞれ発話が自立語1語、非自立語1語で構成されているかを表わす。自立語1語で表現されることの多い対話行為としては「応答（平叙）」や「あいづち」がある。 $f_{28}$ は同じ単語が発話の中で複数回使われているかを表わす。応答表現や「あいづち」によく見られる繰り返しによる強調表現に対応するために導入した。

対話行為を推定する二値分類器を学習する際には、発話の特徴量のベクトルで表現する。特徴量ベクトルの重みは、その特徴量が発話に出現していれば1、それ以外は0とする。

### 3.3 特徴セットの最適化

ここでは、個々の対話行為毎に、対話行為推定のための特徴を最適化する手法について述べる。

#### 3.3.1 最適な特徴セットの決定

個々の対話行為に対し、表2に示した特徴の中から、その対話行為の分類に有効でないものを削除することで、対話行為毎に最適な特徴のセットを決める。そのアルゴリズムを図2に示す。 $E$ は全特徴の集合、 $E'$ は最適化された特徴の集合である。 $f(X)$ は、 $X$ を特徴として学習した分類器の開発データにおけるF値<sup>2</sup>である。特徴 $f_i$ を除いたときのF値 $f(E \setminus \{f_i\})$ が全特

```

Input:  $E = \{f_1, f_2, \dots, f_n\}$ 
Output:  $E'$ 
1: while true do
2:    $E' \leftarrow \emptyset$ 
3:   for all  $f_i \in E$  do
4:     if  $f(E) \geq f(E \setminus \{f_i\})$  then  $E' \leftarrow E' \cup \{f_i\}$ 
5:   end for
6:   if  $E = E'$  then return  $E'$ 
7:   if  $f(E') \geq f(E)$  then
8:      $E \leftarrow E'$ 
9:   else
10:     $f_x = \arg \max_{f_i} f(E \setminus \{f_i\}) - f(E)$ ;  $E \leftarrow E \setminus \{f_x\}$ 
11:   end if
12: end while

```

図2 特徴の選択アルゴリズム

<sup>2</sup> 発話がある対話行為に該当するか否かを判定する二値分類のF値。

徴を用いたときの F 値  $f(E)$  よりも低ければ,  $f_i$  を有効な特徴とみなして  $E'$  に入れ, そうでなければ削除する. これを全ての特徴について行い, 1つ以上の特徴が削除されたら, 残された特徴を新たに全特徴の集合とみなして同様の操作を行う. ただし, 個別に評価したときに有効でない特徴は  $E'$  に残されていないにも関わらず, 7行目の段階で複数の特徴が削除された  $E'$  を用いたときの F 値がもとの  $E$  と比べて低くなることもある. そのときは, 特徴を削除することによって最も F 値が向上する (最も悪影響を与える) ものを 1つ選択し, それのみを削除した特徴の集合を新たな  $E$  とする (10行目). これを特徴が削除されなくなるまで繰り返す.

### 3.3.2 対話行為列の長さの最適化

特徴  $f_{18}$  と  $f_{19}$  は, 「質問 (YesNo)」の次には「応答 (YesNo)」の発話が出現しやすいといったように, 対話行為の並びを考慮するために導入した. しかし, 直前だけでなく, 2つ以前の発話からの対話の流れが対話行為の推定に有効である場合も考えられる. このとき, どれくらい前の発話を辿ればよいか, つまり過去の発話の対話行為列の長さをいくつに設定すればよいかは, 分類対象とする対話行為によって異なると考えられる.

本研究では, 特徴  $f_{18}$  と  $f_{19}$  をそれぞれ相手もしくは話者自身の過去の  $N_h (= 1, 2, 3, 4, 5)$  個の発話の対話行為の列とし,  $N_h$  の値を対話行為に応じて最適化する. すなわち, 対話行為毎に, 開発データでの F 値が最大となる  $N_h$  を選択する. また,  $N_h$  の値が大きいきには特徴量の数が増えるため, 特徴量の選択を行う. 具体的には, 特徴量と対話行為の相関の強さを  $\chi^2$  値で測り, それが閾値  $T_h$  よりも小さい特徴量を削除する.  $T_h$  は 0, 1, 5, 10 のいずれかとし,  $N_h$  と同様に開発データでの F 値が最大となる値を選択することで最適化する.

$N_h$  と  $T_h$  の最適化は, 3.3.1 で述べた最適な特徴セットを決定する前に行う. このとき, 特徴は  $f_1$  (単語 n-gram) と  $f_{18}$  もしくは  $f_{19}$  のみを使用する<sup>3</sup>.

## 3.4 組み合わせ特徴量

本研究で使用する LIBLINEAR では特徴量間の相関関係は考慮されていない. しかし, 特徴量の組み合わせが対話行為の分類に特に有効に働く可能性がある. そのため, 2つの特徴量を組み合わせた特徴量も使用する. 以下, これを「組み合わせ特徴量」と呼ぶ. ただし, 全ての特徴量を組み合わせると特徴量の数が増大するため, 図2のアルゴリズムにより得られたそれぞれの対話行為に最適な特徴セットの F 値と, その特徴セットから1つの特徴を除いた場合の F 値の差が最も大きい特徴を「最も有効な特徴」と定義し, 最も有効な特徴の特徴量とそれ以外の特徴量の組のみを組み合わせ特徴量として導入する.

<sup>3</sup> 予備実験では, 先に最適な特徴のセットを決定し, その後  $N_h$  と  $T_h$  の最適化を行う手法も試したが, 対話行為推定の F 値はわずかに低下した.

### 3.5 対話行為の選択

本項では、個々の対話行為の二値分類器の出力結果から、最も適切な対話行為を1つ選択する手法について述べる。

#### 3.5.1 判定の信頼度による選択

対話行為の二値分類器が出力する信頼度を比較し、それが最も高い対話行為を選択する。具体的には、式(1)にしたがって最終的に選択する対話行為  $\hat{d}$  を決定する。  $r(d_i)$  は対話行為  $d_i$  の判定の信頼度を表わす。

$$\hat{d} = \arg \max_{d_i} r(d_i) \quad (1)$$

#### 3.5.2 信頼度を特徴量とする機械学習による手法

9つの対話行為の二値分類器の出力結果を特徴量とし、対話行為を選択するモデルを機械学習する。当然だが、3.5.1で述べた手法において、信頼度1位の対話行為が常に正解となるわけではない。ここでの狙いは、「対話行為  $d_a$  と  $d_b$  について、 $d_a$  の信頼度が1位であるが、 $d_a$  と  $d_b$  の信頼度の差がそれほど大きくないときは、 $d_b$  が正解である可能性が高い」といった傾向を自動的に学習することにある。この手法では以下の特徴量を用いる。

- 対話行為  $d_i$  の判定の信頼度。
- 信頼度の順位が  $n$  位の対話行為の判定の信頼度。 ( $n = 1, 2, 3$ )

上記の特徴量の重みは信頼度の値とする。後者の特徴量は、テキスト分類において、他クラスの信頼度を考慮する有効性が高橋らにより報告されている (Takahashi, Takamura, and Okumura 2007) ことから設定した。機械学習アルゴリズムとしてロジスティック回帰 (LIBLINEAR) を用いた。

#### 3.5.3 信頼度に対する重み付けに基づく手法

予備実験の結果、「自己開示」以外の対話行為を持つ発話に対して「自己開示」が誤って選択される事例が多いことがわかった。「自己開示」の信頼度は他の対話行為に比べて平均的に高く、「自己開示」が最終的に選ばれやすいためであった。これは、4.1項で後述するように、訓練データにおける「自己開示」の出現頻度が高いためと考えられる。このような信頼度の不均衡を是正するため、式(2)にしたがって対話行為を選択する。

$$\hat{d} = \begin{cases} \arg \max_{d_i} w_i \cdot r(d_i) & \text{if rank}(1)=\text{自己開示} \\ \arg \max_{d_i} r(d_i) & \text{if それ以外} \end{cases} \quad (2)$$

$\text{rank}(1)$  は信頼度の順位が1位の対話行為を表わす。  $w_i$  は対話行為  $d_i$  の信頼度に与える重みであり、「自己開示」以外の対話行為の信頼度を大きくする働きをする。また、「自己開示」に対

する重みは1と設定する.

信頼度の重みを反復推定するアルゴリズムを図3に示す. 変数  $j$  は反復のステップを表わす変数で, 7~13行目の処理を繰り返す. 開発データ  $D_{dev}$  における発話  $u_k$  に対し, その正解の対話行為が自己開示ではなく, 誤って自動推定された対話行為が自己開示であり,  $uncertainty(u_k)$  が閾値  $TU_i$  より大きいとき (9行目), 正解の対話行為  $d_i$  に対する重み  $w_i^{(j)}$  を10行目の式にしたがって更新する.  $uncertainty(u_k)$  は発話  $u_k$  に対する対話行為推定の不確かさを表わす指標であり, 9つの対話行為に対する判定の信頼度  $r(d_i)$  を得たとき, その1位の信頼度と2位の信頼度の比と定義する<sup>4</sup>.  $TU_i$  は対話行為  $d_i$  に対する重みを更新するか否かを定める  $uncertainty(u_k)$  の閾値である. 基本的には, 不正解となった「自己開示」の信頼度と正解の対話行為  $d_i$  の信頼度の差が大きいときほど  $w_i^{(j)}$  により大きい値を加える.  $w_i^{(j)}$  の値を増やすことにより, 正解の対話行為  $d_i$  の信頼度が高くなり, 選ばれる可能性が増す.  $\delta$  は重みの1回当たりの変動量を調整するパラメタである. 本研究では  $\delta = 0.001$  とした. 開発データの全ての発話について重みの調整が終わったら, 新しい重みを用いて, システムによる自動推定の結果を更新する (13行目).

一般に  $w_i^{(j)}$  は収束するが, 本研究では収束後の重みではなく, 1回の反復毎に開発データにおける対話行為推定の改善度  $eval_j(d_i)$  を測り, これが最も高い時点での重みを選択する (15行目).  $eval_j(d_i)$  の定義は式 (3) であり, 対話行為が  $d_i$  である発話のうち重み付けによって新た

```

1:  $gold(u_k) \stackrel{def}{=} \text{発話 } u_k \text{ の正解の対話行為}$ 
2:  $predict_j(u_k) \stackrel{def}{=} j \text{ 回目の反復が終わった時点で自動推定された } u_k \text{ の対話行為}$ 
3:  $w_i^{(j)} \stackrel{def}{=} j \text{ 回目の反復における対話行為 } d_i \text{ の重み}$ 
4:  $r'_j(d_i) \stackrel{def}{=} w_i^{(j)} \cdot r(d_i)$  # 重み付けによって調整された対話行為  $d_i$  の信頼度
5:  $\forall i \ w_i^{(0)} \leftarrow 1$  # 初期化
6: for  $j = 1$  to 500 do
7:    $\forall i \ w_i^{(j)} \leftarrow w_i^{(j-1)}$ 
8:   for all  $u_k \in D_{dev}$  do
9:     if  $gold(u_k) = d_i$  and  $d_i \neq \text{自己開示}$  and  $predict_{j-1}(u_k) = \text{自己開示}$  and
        $uncertainty(u_k) > TU_i$  then
10:        $w_i^{(j)} \leftarrow w_i^{(j)} + \delta \times \left( \frac{r'_{j-1}(\text{自己開示}) - r'_{j-1}(d_i)}{r'_{j-1}(\text{自己開示})} \right)$ 
11:     end if
12:   end for
13:    $update(predict_j)$ 
14: end for
15:  $\forall i \ w_i \leftarrow w_i^{(j)}$  where  $j = \arg \max_j eval_j(d_i)$ 
16: return  $\{w_i\}$ 

```

図3 信頼度に対する重みを決定するアルゴリズム

<sup>4</sup> 1位と2位の信頼度が近ければ近いほど, 1位の対話行為が正しくない可能性が高い.

に正解となった発話数 ( $|B|$ ) と, 対話行為が「自己開示」である発話のうち重み付けによって新たに不正解となった発話数 ( $|W|$ ) の差である<sup>5</sup>.

$$eval_j(d_i) = |B| - |W| \quad (3)$$

$$B = \{u_k \mid gold(u_k) = d_i \wedge predict_0(u_k) \neq gold(u_k) \wedge predict_j(u_k) = gold(u_k)\}$$

$$W = \{u_k \mid gold(u_k) = \text{自己開示} \wedge predict_0(u_k) = gold(u_k) \wedge predict_j(u_k) \neq gold(u_k)\}$$

本手法では,  $uncertainty(u_k)$  が低いときは重みの更新を行わない. これは個々の対話行為の二値分類器の結果が十分に信頼できるとみなしているためである. 閾値  $TU_i$  は重みの更新を行うか行わないかをコントロールする働きをする.  $TU_i$  は重み  $w_i$  の推定に用いたものとは別の開発データを用いて最適化する.  $TU_i$  を変動させ, 学習した重みを用いたシステムの  $eval$  の値が最大となる閾値を選択する.

### 3.5.4 特定の対話行為の組に対して機械学習で識別する手法

対話行為の中には互いに識別が難しい組み合わせがある. 表 3 は, 対話行為のそれぞれの組に対し, 一方の対話行為の信頼度の順位が 1 位でかつ不正解, もう一方の対話行為の信頼度の順位が 2 位でかつ正解となる発話の開発データにおける数を示している. この表において発話数 (誤り数) の多い対話行為の組は, 特に判定が難しいと考えられる. ここでは, このような対話行為の組に対し, 適切な対話行為を選択する分類器を機械学習することを試みる. ただし, 「自己開示」( $d_1$ ) については, 3.5.3 で述べた信頼度の重み付けによる手法で対応することとし, ここでは  $d_1$  を含まない組の中で表 3 における誤り発話数が多い組に着目する. 具体的には, 他と比べて誤り発話数の多い (あいづち, フィラー) と (質問 (YesNo), 確認) の 2 つの組につい

表 3 信頼度 1 位が不正解, 2 位が正解となる対話行為の組と発話数

	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$	$d_7$	$d_8$	$d_9$
$d_1$		42	72	1	231	161	111	118	54
$d_2$			36	0	4	1	0	102	3
$d_3$				0	5	8	4	1	1
$d_4$					6	6	8	0	0
$d_5$						2	4	6	7
$d_6$							151	9	0
$d_7$								0	0
$d_8$									0

$d_1$ : 自己開示,  $d_2$ : 質問 (YesNo),  $d_3$ : 質問 (What),  $d_4$ : 応答 (YesNo),  $d_5$ : 応答 (平叙),  $d_6$ : あいづち,  $d_7$ : フィラー,  $d_8$ : 確認,  $d_9$ : 要求

<sup>5</sup>  $predict_0(u_k)$  は重み付けしない手法で選択された発話  $u_k$  の対話行為を表わす.

て、機械学習により適切な対話行為を選択する。以上をまとめると、本手法は式(4)にしたがって  $\hat{d}$  を決定する。

$$\hat{d} = \begin{cases} \arg \max_{d_i} w_i \cdot r(d_i) & \text{if rank(1)=}d_1(\text{自己開示}) \\ \text{classify}(\text{rank(1),rank(2)}) & \text{if } \{\text{rank(1),rank(2)}\} = \{d_6, d_7\} \text{ or } \{d_2, d_8\} \\ \arg \max_{d_i} r(d_i) & \text{if それ以外} \end{cases} \quad (4)$$

rank(1), rank(2) は判定の信頼度が 1 位, 2 位の対話行為を表わし,  $\text{classify}(x, y)$  は 2 つの対話行為  $x, y$  の中から一方を選択する分類器である。  $\text{classify}(x, y)$  の学習に使う特徴量は, 組み合わせ特徴量も含めて対話行為  $x$  と  $y$  の分類に用いる特徴量の和集合とし, 学習には LIBLINEAR を用いる。

## 4 評価実験

### 4.1 データ

対話コーパスとして, 人間同士の自由対話を書き起こした名大会話コーパス (国立国語研究所 2001) を用いた。実験では, 対話コーパスの中から参加者が二名の対話のみを選択し, 各発話に対し対話行為タグを人手で付与した。対話数は 97, 発話数は 91,906 である。3 対話について二者によって対話行為タグを付与したところ, 一致率は 77.3%,  $\kappa$  係数は 0.636 であった。コーパスにおける対話行為の出現頻度ならびに割合を表 4 に示す。  $d_1$  (自己開示) が最も多く, 全体の 6 割弱を占めている。一方,  $d_9$  (要求) は最も少なく, それが占める割合は 1%未満である。コーパスをおよそ 80%, 10%, 10%に分割し, 77 対話 (74,228 発話) を訓練データ, 10 対話 (8,984 発話) を開発データ, 10 対話 (8,694 発話) をテストデータとした。開発データは最適な特徴の選択やパラメタの最適化のために用いた。それぞれのデータにおける対話行為の頻度分布は全体とほぼ同じであった。

### 4.2 パラメータ最適化

特徴  $f_{18}$  (相手の過去の発話の対話行為列),  $f_{19}$  (話者の過去の発話の対話行為列) について, 3.3.2 で述べたように, 過去の対話行為列の長さ  $N_h$  ならびに特徴量選択の閾値  $T_h$  の最適化を行った。

表 4 実験データにおける対話行為の出現頻度の分布

$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$	$d_7$	$d_8$	$d_9$
53,625	6,423	3,943	2,123	7,492	9,217	4,404	3,930	749
(58.3%)	(7.0%)	(4.3%)	(2.3%)	(8.2%)	(10.0%)	(4.8%)	(4.3%)	(0.8%)

本実験では、テストデータにおける発話の対話行為を推定する際、 $f_{18}$  と  $f_{19}$  の特徴量は正解の対話行為を用いる。実際には、過去の発話の対話行為は自動的に推定すべきである。しかし、このような実験設定では、対話行為推定の誤りが次の発話の対話行為の推定に影響し、対話行為の誤推定が前の発話の対話行為の誤りによるものか、それとも提案手法の不備など他の要因によるものなのかを区別できない。今回の実験では、提案手法の有効性を確認することに重点を置き、過去の発話の対話行為の分類に誤りはないという理想的な条件下で実験を行った。開発データにおける F 値が最大となった  $N_h$  と  $T_h$  の値を表 5 に示す。この表ではパラメタの最適化を行わないとき ( $N_h = 1, T_h = 0$ ) の F 値も示した。‘—’ は  $N_h = 1, T_h = 0$  のときに F 値が最大になった場合、すなわちパラメタの最適化によって F 値が向上しなかった場合を表わす。

この結果から、対話行為毎に話者自身の過去の発話の対話行為列、相手の過去の発話の対話行為列の最適な長さが異なることが示された。特に、「フィラー」については F 値が 11 もしくは 14 ポイント向上しており、パラメタ最適化の影響が大きい。これは「フィラー」の発話を認識するためにはそれまでの対話の流れが重要な情報であることを示唆する。一方で、「質問 (YesNo)」、「応答 (YesNo)」については自身の過去の発話、相手の過去の発話ともに  $N_h = 1, T_h = 0$  のときが最良となっている。「質問 (YesNo)」については、前の発話の対話行為の影響が小さいと考えられるので、対話行為列の長さを変化させても影響がなかったと考えられる。「応答 (YesNo)」については、前の相手の最後の発話が「質問」であることが多いため、 $f_{18}$  についてはひとつ前の相手の発話の対話行為だけを特徴量とすれば十分と考えられる。一方、 $f_{19}$  については、F 値が他の対話行為と比べて極端に低い。 $N_h, T_h$  の最適化の際には  $f_1$  (単語 n-gram) のみの特徴としていることが原因と考えられる。

表 5 過去の発話の対話行為の特徴のパラメタ

対話行為	$f_{18}$ (相手)			$f_{19}$ (話者)			対話行為	$f_{18}$ (相手)			$f_{19}$ (話者)		
	$N_h$	$T_h$	F 値	$N_h$	$T_h$	F 値		$N_h$	$T_h$	F 値	$N_h$	$T_h$	F 値
自己開示	1	0	0.856	1	0	0.851	あいづち	1	0	0.637	1	0	0.588
	—	—	—	5	10	0.851		2	10	0.651	3	1	0.604
質問 (YesNo)	1	0	0.722	1	0	0.737	フィラー	1	0	0.278	1	0	0.355
	—	—	—	—	—	—		2	5	0.390	4	5	0.496
質問 (What)	1	0	0.714	1	0	0.707	確認	1	0	0.333	1	0	0.357
	—	—	—	2	0	0.708		5	0	0.342	—	—	—
応答 (YesNo)	1	0	0.771	1	0	0.033	要求	1	0	0.405	1	0	0.395
	—	—	—	—	—	—		2	5	0.411	3	0	0.410
応答 (平叙)	1	0	0.467	1	0	0.205							
	2	0	0.483	—	—	—							

### 4.3 特徴セットの最適化の結果

個々の対話行為に対して選択された特徴を表6に示す。表6の結果から、対話行為毎に有効な特徴が大きく異なることが確認された。 $f_1$  (単語 n-gram) は全ての対話行為に共通して有効な特徴である。一方で、 $f_2$  (前発話の単語 n-gram) や  $f_8$  (前発話の文末付属語列) は全ての対話行為で不要であり、前の相手の発話の内容は有効な特徴ではないと考えられる。表7は3.4項で定義した最も有効な特徴の一覧である。これらも対話行為毎に異なるが、 $f_1, f_{18}, f_{19}$  のいずれかが選ばれており、これらが特に重要な特徴であることがわかる。

### 4.4 信頼度の重みの推定

3.5.3で述べた手法において、対話行為毎の信頼度の重み  $w_i$  は開発データを用いて推定した。一方、閾値  $TU_i$  は、開発データとは別のデータで最適化する必要がある。本実験では、訓練データの8分割交差検定により  $TU_i$  を最適化した。交差検定の際には、機械学習の特徴や重み  $w_i$  は開発データで決定したものをを用いるが、分類器の学習は分割されたデータ毎にやり直した。 $TU_i$

表6 選択された特徴

	$f_1$	$f_2$	$f_3$	$f_4$	$f_5$	$f_6$	$f_7$	$f_8$	$f_9$	$f_{10}$	$f_{11}$	$f_{12}$	$f_{13}$	$f_{14}$	$f_{15}$	$f_{16}$	$f_{17}$	$f_{18}$	$f_{19}$	$f_{20}$	$f_{21}$	$f_{22}$	$f_{23}$	$f_{24}$	$f_{25}$	$f_{26}$	$f_{27}$	$f_{28}$
$d_1$	✓				✓		✓		✓				✓	✓				✓	✓	✓	✓		✓	✓	✓		✓	✓
$d_2$	✓				✓		✓				✓	✓		✓	✓		✓		✓		✓	✓				✓	✓	✓
$d_3$	✓		✓		✓						✓		✓	✓	✓	✓	✓			✓	✓	✓		✓				✓
$d_4$	✓		✓				✓						✓			✓	✓	✓			✓							✓
$d_5$	✓					✓			✓	✓	✓	✓				✓		✓	✓	✓		✓	✓	✓				✓
$d_6$	✓		✓		✓				✓	✓	✓				✓			✓	✓	✓	✓		✓		✓	✓		✓
$d_7$	✓			✓	✓	✓				✓		✓	✓			✓	✓	✓	✓	✓		✓	✓			✓		✓
$d_8$	✓		✓		✓	✓			✓						✓					✓				✓		✓		✓
$d_9$	✓			✓	✓	✓	✓		✓		✓	✓	✓	✓	✓	✓	✓		✓		✓	✓		✓	✓	✓	✓	✓

$d_1$ : 自己開示,  $d_2$ : 質問 (YesNo),  $d_3$ : 質問 (What),  $d_4$ : 応答 (YesNo),  $d_5$ : 応答 (平叙),  $d_6$ : あいづち,  $d_7$ : フィラー,  $d_8$ : 確認,  $d_9$ : 要求

表7 対話行為の分類に最も有効な特徴

自己開示	$f_{18}$ : 相手の過去の発話の対話行為
質問 (YesNo)	$f_1$ : 単語 n-gram
質問 (What)	$f_1$ : 単語 n-gram
応答 (YesNo)	$f_{18}$ : 相手の過去の発話の対話行為
応答 (平叙)	$f_{19}$ : 話者の過去の発話の対話行為
あいづち	$f_{18}$ : 相手の過去の発話の対話行為
フィラー	$f_{19}$ : 話者の過去の発話の対話行為
確認	$f_1$ : 単語 n-gram
要求	$f_1$ : 単語 n-gram

表 8 対話行為ごとの  $w_i$  と  $TU_i$ 

	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$	$d_7$	$d_8$	$d_9$
$w_i$	1.123	1.437	1.000	1.207	1.000	1.630	1.000	1.537
$TU_i$	0.4	0.5	—	0.5	—	0.5	—	0.5

表 9 8 分割交差検定における分割データ毎の eval 値

	$TR_1$	$TR_2$	$TR_3$	$TR_4$	$TR_5$	$TR_6$	$TR_7$	$TR_8$	合計
$d_2$ : 質問 (YesNo) ( $TU_2 = 0.4$ )	1	-1	-2	1	1	2	1	2	5
$d_5$ : 応答 (平叙) ( $TU_5 = 0.5$ )	5	1	10	8	6	9	0	4	43
$d_7$ : フィラー ( $TU_7 = 0.5$ )	8	12	15	8	4	4	4	-6	49

を 0 から 0.9 まで 0.1 刻みで変動させ、式 (3) の eval の値が一番大きい閾値を選択した。「自己開示」以外の対話行為に対する  $w_i$  と  $TU_i$  の一覧を表 8 に示す。 $d_4$ (応答 (YesNo)),  $d_7$ (あいづち),  $d_8$ (確認) の 3 つの対話行為については、信頼度に対する重み付けを行っても対話行為推定結果は向上しなかったため、重みを 1 に設定している。すなわち、これらの対話行為の信頼度に対しては重み付けを行わない。

表 9 は、例として  $d_2$ ,  $d_5$ ,  $d_7$  の 3 つの対話行為について、8 分割交差検定において分割された個々のデータ ( $TR_1 \sim TR_8$ ) に対する eval の値を示している<sup>6</sup>。eval の値は対話データによってはばらつきが見られ、負の値になる(重み付けによって悪化する)こともある。この結果から、信頼度に対する重み付けに基づく手法は、対話によって効果的に働く場合とそうでない場合があることがわかった。信頼度に対する重み付けは、「自己開示」の判定の信頼度が他の対話行為に比べて高いことを是正するための手法であるが、自己開示の発話の出現のしやすさは対話の内容に強く依存しており、自己開示の発話が多く出現する対話に対しては、「自己開示」以外の対話行為を選択しやすくする本手法が有効に働かなかったと推察できる。

#### 4.5 対話行為推定の評価

対話行為を推定する提案手法の性能を評価する。評価基準は、各対話行為の推定の精度、再現率、F 値、ならびにこれら 3 つの全対話行為についてのマクロ平均とマイクロ平均である。なお、精度および再現率のマイクロ平均は正解率(システムが選択した対話行為と正解の対話行為が一致する割合)に等しい。提案手法を 2 つのベースラインと比較する。一つは、全ての特徴を用いて、9 つの対話行為のいずれかを選択する分類器を LIBLINEAR で学習する手法 ( $BL_a$ ) である。もう一つは、3.3 項で説明した方法で特徴を選択する手法 ( $BL_s$ ) である。提案手法が個々の対話行為毎に最適な特徴を選択するのに対し、 $BL_s$  では特徴のセットを 1 つだけ選択し、

<sup>6</sup>  $TU_i$  は表 9 における「合計」が最も大きい値を選んで最適化している。

それを用いて全ての対話行為を分類する。一方、提案手法として、3.5.1で述べた信頼度を比較する手法 ( $Pro_p$ )、3.5.2で述べた信頼度を特徴量とした機械学習を用いる手法 ( $Pro_m$ )、3.5.3で述べた「自己開示」以外の対話行為の信頼度に対して高い重みを与える手法 ( $Pro_w$ )、3.5.4で述べた判定の難しい対話行為の組に対して機械学習で適切な対話行為を選択する手法 ( $Pro_b$ ) の4つを評価する。

まず、発話がある対話行為を持つか否かを判定するタスク（以下、「個別対話行為判定タスク」と呼ぶ）についてベースラインと提案手法を比較する。言い換えれば、個別対話行為判定タスクでは、図1の第1段階における対話行為毎に学習した分類器の性能を評価する。表10は同タスクにおける  $BL_s$  と  $Pro_p$  の精度 (P)、再現率 (R)、F 値 (F) を示している。表10(a)は開発データの結果であり、対話行為毎に特徴を最適化することによって、全ての対話行為について評価値が同等もしくは向上していることが確認できる。一方、表10(b)はテストデータの結果であり、 $Pro_p$  は  $BL_s$  に比べて F 値のマクロ平均が 2.9 ポイント向上した。しかしながら、「あいづち」と「要求」については F 値が低下している。これは開発データとテストデータとで対話の内容が異なるため、両データにおいて最適な特徴が一致していないためと考えられる。この結果は、自由対話では様々なトピックが話題に挙がるため、対話行為分類のための最適な特徴を実験的に決定することが難しいことを示唆する。

表11は、発話に対して9つの対話行為の中から該当するものを推定するタスク（以下、「対話行為推定タスク」と呼ぶ）における各手法の評価値を示している。2つのベースラインを比較すると、 $BL_s$  はマクロ平均では  $BL_a$  を上回るが、マイクロ平均は等しい。対話行為を区別せずに単純に特徴を最適化しても、正解率は向上しないことがわかる。一方、4つの提案手法の F 値のマイクロ平均はいずれもベースラインよりも高い。最も結果が良かったのは手法  $Pro_b$  であった。 $BL_s$  と  $Pro_b$  の結果をマクネマー検定で検定したところ、5%の有意水準で有意差が

表 10 個別対話行為判定タスクの結果

	(a) 開発データ						(b) テストデータ					
	$BL_s$			$Pro_p$			$BL_s$			$Pro_p$		
	P	R	F	P	R	F	P	R	F	P	R	F
自己開示	.851	.907	.878	.855	.920	.886	.848	.919	.881	.856	.925	.889
質問 (YesNo)	.699	.734	.716	.732	.751	.742	.630	.838	.719	.763	.680	.719
質問 (What)	.820	.590	.687	.809	.651	.721	.327	.919	.483	.787	.672	.725
応答 (YesNo)	.902	.847	.874	.951	.875	.911	.827	.871	.848	.872	.885	.879
応答 (平叙)	.737	.687	.711	.760	.748	.754	.804	.741	.771	.804	.798	.801
あいづち	.713	.591	.647	.763	.609	.678	.776	.731	.753	.763	.717	.739
フィラー	.652	.359	.463	.699	.440	.540	.619	.311	.414	.612	.356	.450
確認	.566	.236	.333	.644	.279	.389	.191	.819	.310	.680	.254	.370
要求	.750	.207	.324	.850	.293	.436	.618	.347	.444	.714	.204	.317
マクロ平均	.743	.573	.626	.785	.618	.673	.627	.722	.625	.761	.610	.654

表 11 対話行為推定タスクの結果

	$BL_a$			$BL_s$			$Pro_p$			$Pro_m$			$Pro_w$			$Pro_b$		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
$d_1$	.855	.949	.899	.851	.951	.898	.852	.953	.900	.858	.951	.902	.859	.949	.901	.859	.949	.901
$d_2$	.743	.742	.742	.762	.745	.753	.754	.751	.752	.755	.761	.758	.754	.753	.753	.760	.753	.756
$d_3$	.739	.667	.701	.787	.672	.725	.807	.689	.743	.799	.700	.746	.797	.706	.749	.797	.706	.749
$d_4$	.877	.885	.881	.874	.900	.887	.876	.880	.878	.889	.885	.887	.876	.880	.878	.876	.880	.878
$d_5$	.820	.804	.812	.819	.772	.795	.818	.812	.815	.805	.846	.825	.811	.839	.824	.811	.839	.824
$d_6$	.751	.751	.751	.768	.730	.748	.758	.724	.741	.763	.716	.738	.758	.724	.741	.790	.699	.741
$d_7$	.660	.378	.481	.608	.412	.491	.607	.399	.482	.593	.423	.494	.598	.423	.495	.627	.553	.588
$d_8$	.658	.289	.402	.634	.318	.424	.678	.265	.381	.709	.258	.379	.678	.265	.381	.687	.276	.394
$d_9$	.808	.214	.339	.724	.214	.331	.773	.173	.283	.680	.173	.276	.643	.184	.286	.643	.184	.286
Ma	.768	.631	.667	.759	.635	.672	.769	.628	.664	.761	.635	.667	.753	.636	.668	.761	.649	.680
Mi	.819	.819	.819	.819	.819	.819	.821	.821	.821	.823	.823	.823	.824	.824	.824	.825	.825	.825

$d_1$ : 自己開示,  $d_2$ : 質問 (YesNo),  $d_3$ : 質問 (What),  $d_4$ : 応答 (YesNo),  $d_5$ : 応答 (平叙),  $d_6$ : あいづち,  $d_7$ : フィラー,  $d_8$ : 確認,  $d_9$ : 要求, Ma: マクロ平均, Mi: マイクロ平均

あった.  $Pro_b$  が選択した対話行為と正解の対話行為の対応表を付録 A に示す.

対話行為毎に結果を比較すると, 「応答 (YesNo)」「あいづち」「確認」「要求」については  $Pro_b$  は  $BL_s$  に比べて F 値は改善しなかったが, 「自己開示」「質問 (YesNo)」「質問 (What)」「応答 (平叙)」「フィラー」については F 値が 0.3~9.7 ポイント改善した.

$Pro_p$  と  $Pro_m$  を比較すると,  $Pro_m$  は「あいづち」「確認」「要求」以外の対話行為でより高い F 値が得られており, 信頼度を特徴量とした機械学習の手法が有効であることを示している. 「自己開示」について  $Pro_p$  と  $Pro_w$  を比較すると, 再現率は  $Pro_p$  の方が高いが, 精度ならびに F 値では  $Pro_w$  が上回る. 信頼度に重み付けを行う  $Pro_w$  は, 判定の信頼度が全般に高い「自己開示」が過度に選ばれることを抑制するための手法であるが, この手法により「自己開示」の false positive の誤りが減少したことが確認された. また, 表 8 で重みを 1 より大きく設定した全ての対話行為で F 値が向上した.  $Pro_b$  は  $Pro_w$  と比べて, 誤りが多かったために改めて機械学習で分類し直した「質問 (YesNo)」「フィラー」「確認」の結果が改善されていることが確認できた. ただし, 「あいづち」については, 精度, 再現率に変化はあったが, F 値は変化しなかった.

本論文では, ベースラインで精度や再現率が低い対話行為に対して推定の性能を向上させることを目指したが, 一部の対話行為についてはその目標が達成されていない. 具体的には, ベースラインで性能の低い「フィラー」の評価値は向上しているが, 「確認」や「要求」については逆にベースラインよりも低くなっている. 「確認」については, 表 10 より, 個別対話行為分類タスクでは提案手法はベースラインを上回っているため, 図 1 における二段階の処理のうち, 第 1 段階で「確認」に該当するかを判定する時点では性能の向上が見られるものの, 第 2 段階

表 12 組み合わせ特徴量の評価

	$BL_a$	$BL_s$	$Pro_p$	$Pro_m$	$Pro_w$	$Pro_b$
組み合わせ特徴量なし	.808	.815	.816	.816	.819	.823
組み合わせ特徴量あり	.819	.819	.821	.823	.824	.825

の対話行為を推定する段階で誤りを多く生じていることがわかる。一方、「要求」については表 10 でも表 11 でも提案手法はベースラインより劣る。この原因として、コーパスにおいて「要求」の対話行為を持つ発話の数が他の対話行為と比べて極端に少ないことが考えられる。

組み合わせ特徴量の有効性を評価するために、組み合わせ特徴量を使用したモデルと使用しないモデルの F 値のマイクロ平均（正解率）を比較した。結果を表 12 に示す。いずれの手法も組み合わせ特徴量を用いることで F 値が向上していることから、組み合わせ特徴量の有効性が確認できた。

#### 4.6 機械学習アルゴリズムの比較

前項までの実験では機械学習アルゴリズムとして L2 正則化ロジスティック回帰を用いたが、本項ではこれと他の機械学習アルゴリズムを比較する。また、対話行為毎に適切な特徴セットを設定するという提案手法の基本的な考え方が他の機械学習アルゴリズムでも有効であるかを検証する。そのため、 $BL_a$ （全特徴を用いたベースライン）、 $BL_s$ （対話行為を区別しないで特徴を選択したベースライン）、ならびに提案手法のうち最も基本的な  $Pro_p$  を比較する実験を行う。

比較する機械学習アルゴリズムは SVM とする。カーネル関数として、線形カーネル、多項式カーネル（カーネルの次数は 3）、Radial Basis Function (RBF) カーネル、シグモイドカーネルの 4 つを用いる。 $Pro_p$  では、対話行為毎に特徴を選択するために、特徴セットを変えて学習とテストを繰り返す必要があるが、SVM の学習は非常に時間がかかるため、現実的な時間では特徴選択が終了しない。そこで、高速な LIBLINEAR を用いて選択された特徴のセット（表 6）を用い、対話行為毎の二値分類器を学習するときのみ SVM を用いる。同様に、 $BL_s$  も LIBLINEAR を用いて選択された特徴のセットを用いる。また、L2 正則化ロジスティック回帰とは異なり多項式カーネルの SVM では特徴量の組み合わせも学習時に考慮されるため、ここでの実験では組み合わせ特徴量はいらない。SVM の学習には LIBSVM<sup>7</sup>を用いる。学習パラメータはデフォルト値を用いる。 $Pro_p$  で用いる個々の対話行為判定の信頼度は、LIBSVM が出力する確率とする。

SVM による対話行為推定の F 値のマイクロ平均（正解率）を表 13 に示す。比較のため、L2 正則化ロジスティック回帰を用いたときの結果（表 12 の組み合わせ特徴量なしの結果）も再掲する。\*と † はマクネマー検定で  $Pro_p$  と他の手法の差を検定した結果を表わす。\*はロジス

<sup>7</sup> <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

表 13 機械学習アルゴリズムの比較

	ロジスティック回帰	SVM (線形)	SVM (多項式)	SVM (RBF)	SVM (シグモイド)
$BL_a$	.808	.805	.560	.763	.763
$BL_s$	.815	.806	.560	.764	.763
$Pro_p$	.816	.801*	.560*	.773*†	.770*

\*:  $p < 0.05$  (vs. ロジスティック回帰の  $Pro_p$ ) , †:  $p < 0.05$  (vs.  $BL_s$ )

ティック回帰の  $Pro_p$  との間に有意水準 5% で有意差が, † は同じ学習アルゴリズムの  $BL_s$  との間に有意差があることを示している. 多項式カーネルの SVM では全ての発話に対して「自己開示」が選択された. これは過学習のためと考えられる.

異なる機械学習アルゴリズムの  $Pro_p$  の結果を比較すると, L2 正則化ロジスティック回帰は全ての SVM よりも正解率が有意に高い. 線形カーネルの SVM については,  $Pro_p$  よりも  $BL_s$  の方が正解率が高いが, ロジスティック回帰の  $Pro_p$  よりは劣る. ただし, 今回の実験では, 特徴選択の際に用いた機械学習アルゴリズム (ロジスティック回帰) と対話行為推定の分類器の学習に用いたアルゴリズム (SVM) が異なる. 特徴選択も SVM で行えば, SVM での分類に適した特徴セットが選ばれて, 正解率が向上する可能性がある. SVM の中で最も結果が良かった線形カーネルの  $BL_s$  については, 特徴選択も線形カーネルの SVM を用いてモデルを学習する追加実験を行ったところ, F 値は 0.806 と変化しなかった<sup>8</sup>. このモデルと提案手法 (ロジスティック回帰の  $Pro_p$ ) とは 5% の有意水準で有意差があった. 他のカーネルの  $BL_s$  や  $Pro_p$  についても特徴選択の段階から SVM を用いるべきであるが, LIBLINEAR 以外では特徴選択に非常に時間がかかるという問題がある. 例えば, Intel Xeon 2.93 GHz, メモリ 8 GB のサーバを用いて, 対話行為「自己開示」に該当するかを判定する二値分類器の学習に, LIBLINEAR では 16 秒を要したのに対し, LIBSVM ではおよそ 168 倍の 2,697 秒を要した. 今回の実験結果からは, LIBLINEAR を用いる手法が対話行為推定の F 値ならびに計算時間の両方の観点から最も優れているといえる.

機械学習アルゴリズム毎に  $BL_s$  と  $Pro_p$  を比較すると, ロジスティック回帰, RBF カーネルの SVM, シグモイド関数の SVM において, 差はそれほど顕著ではないものの,  $Pro_p$  は  $BL_s$  よりも正解率が高かった. 有意な分類器が学習できなかった多項式カーネルの SVM を除けば, 4 つのうち 3 つの機械学習アルゴリズムについて, 対話行為毎に最適な特徴を選択するというアプローチは有効と言える. 但し, 提案手法のアプローチの妥当性をより正確に検証するためには, 異なる対話行為のセットを用いた実験や, 異なる対話コーパスを用いた実験などを

<sup>8</sup> 選択された特徴は異なる. ロジスティック回帰を用いて選択された特徴は  $f_1, f_5, f_8, f_{14}, f_{15}, f_{18}, f_{19}, f_{20}, f_{22}, f_{24}$  であったのに対し, 線形カーネルの SVM で選択された特徴は  $f_1, f_5, f_6, f_8, f_{10}, f_{11}, f_{12}, f_{13}, f_{16}, f_{17}, f_{18}, f_{19}, f_{20}, f_{21}, f_{23}, f_{24}, f_{25}, f_{26}, f_{27}, f_{28}$  であった.

表 14 CRF, ランダムフォレストの結果

$CRF_{all}$		$CRF_{seq}$		$RF$	
$BL_a$	$BL_s$	$BL_a$	$BL_s$	$BL_a$	$BL_s$
.825	.828	.810	.809	.763	.774

行う必要があるだろう。

次に、多値分類の機械学習アルゴリズムである CRF ならびにランダムフォレスト (Breiman 2001) と提案手法を比較する。CRF で分類されるのは系列 (本論文の場合は発話列) であることに注意していただきたい。ここでは、対話全体の発話列を入力として与える手法 ( $CRF_{all}$ ) と、対話の先頭から解析対象となる発話までの発話列を逐次的に入力として与える手法 ( $CRF_{seq}$ ) の 2 つを評価する。実際に対話システムでの利用を想定しているのは  $CRF_{seq}$  である。 $CRF_{all}$  は対話が全て終了するまで対話行為を分類できないため、実際の対話システムに組み込むことは不可能だが、文献 (磯村 他 2009) のようにコーパスへの対話行為のタグ付けを目的とする場合には利用できる。CRF の学習には CRFsuite<sup>9</sup> を、ランダムフォレストの学習には scikit-learn<sup>10</sup> を用いる。

$CRF_{all}$ ,  $CRF_{seq}$ ,  $RF$  (ランダムフォレスト) のそれぞれについて、 $BL_a$  と同じ特徴セット (28 個の全ての特徴) を用いたときと、 $BL_s$  と同じ特徴セットを用いたときの対話行為推定の F 値を表 14 に示す。最良の提案手法である  $Pro_b$  (F 値 0.825) は、 $CRF_{seq}$  と  $RF$  を上回り、マクネマー検定で 5% の有意水準で有意差があることを確認した。 $CRF_{all}$  は  $CRF_{seq}$  よりも解析に利用できる発話数が多いため、F 値が高い。 $BL_s$  と同じ特徴セットを用いたときの  $CRF_{all}$  は  $Pro_b$  を上回るが、 $CRF_{all}$  は対話システムでの利用を想定したモデルではない。

## 5 おわりに

本論文では、自由対話における発話の対話行為を自動推定する新しい手法を提案した。提案手法は、個々の対話行為毎に発話がその対話行為に該当するかを判定する第 1 段階と、第 1 段階の結果から最終的に最も適切な対話行為を選択する第 2 段階から構成される。第 1 段階において、教師あり機械学習のために有効な特徴は対話行為毎に異なるという仮定の下、対話行為毎に最適な特徴のセットを自動的に決定する点に特長がある。評価実験の結果、対話行為を区別せずに特徴の選択を行う手法と比べて、提案手法の対話行為推定の F 値は 0.6 ポイント高かった。F 値の差はそれほど大きくはないものの、統計的に有意な差があることを確認した。

<sup>9</sup> <http://www.chokkan.org/software/crfsuite/>

<sup>10</sup> <http://scikit-learn.org/>

本論文の貢献は以下の通りである。表6に示すように、有効な特徴のセットは対話行為によって異なることを実験的に確認し、対話行為毎に特徴の最適化を行う提案手法のアプローチが有望であることを示した。表13に示したロジスティック回帰ならびに4種のカーネル関数のSVMを用いた検証実験では、ロジスティック回帰、SVM (RBF)、SVM (シグモイド) について、対話行為毎に特徴の最適化を行うことでF値の向上が見られた。ただし、統計的に有意差が確認されたのはSVM (RBF) のみであった。また、過去の対話行為を特徴とするとき、その最適な長さは対話行為毎に異なることを確認した。さらに、提案手法の第2段階において、分類の信頼度を単純に比較して対話行為を1つ選択すると、分類の信頼度が対話行為によって大きく差があるために特定の対話行為（具体的には「自己開示」）が選ばれやすいという問題に対し、適切な対話行為を選択する3つの手法を提案し、それらがF値の向上に貢献することを確認した。一方、対話行為によっては、特徴の最適化により、開発データではF値が向上するもののテストデータは低下することがわかった。自由対話システムでは様々なトピックが話題になることから、対話によって有効な特徴が異なる可能性があり、特徴の最適化を実験的に行う提案手法の問題点も明らかにした。表9に示したように、信頼度の重み付けに基づく手法が対話によって有効に働く場合とそうでない場合があることがわかったが、これも自由対話システムにおけるトピックの多様性に起因すると考えられる。

今後の課題としては、F値が依然として低い「フィルター」「確認」「要求」に対して対話行為推定の性能を向上させることが挙げられる。これらの対話行為の推定に有効な新たな特徴を発見したり、提案手法の第2段階における対話行為選択手法を洗練する必要がある。また、上記の考察を踏まえ、領域適応の技術を応用し、対話の内容が訓練データ・開発データとテストデータとで異なる場合でもF値を低下させない方法を探究することも重要な課題である。4.6項で述べたように、対話行為毎に適切な特徴のセットを設定するという提案手法のアプローチの妥当性を検証するためには更なる実験が必要である。また、自然言語処理分野でも近年盛んに利用されるようになった深層学習 (Kim 2014) との比較も重要である。

## 謝 辞

本論文の査読にあたり、査読者の方から数多くの有益なコメントをいただきました。深く感謝いたします。

## 参考文献

- Breiman, L. (2001). "Random Forests." *Machine Learning*, **45** (1), pp. 5–32.  
Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). "LIBLINEAR:

- A Library for Large Linear Classification.” *The Journal of Machine Learning Research*, **9**, pp. 1871–1874.
- Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., and Matsuo, Y. (2014). “Towards an Open-domain Conversational System Fully Based on Natural Language Processing.” In *Proceedings of COLING 2014*, pp. 928–939.
- 平野徹, 小林のぞみ, 東中竜一郎, 牧野俊朗, 松尾義博 (2016). パーソナライズ可能な対話システムのためのユーザ情報抽出. 人工知能学会論文誌, **31** (1), pp. DSF–B.1–10.
- Inui, N., Ebe, T., Indurkha, B., and Kotani, Y. (2001). “A Case-Based Natural Language Dialogue System Using Dialogue Act.” In *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, pp. 193–198.
- 磯村直樹, 鳥海不二夫, 石井健一郎 (2009). 対話エージェント評価におけるタグ付与の自動化. 電子情報通信学会論文誌. A, 基礎・境界, **92** (11), pp. 795–805.
- Jurafsky, D., Shriberg, E., and Biasca, D. (1997). “Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual.” Tech. rep., Institute of Cognitive Science Technical Report.
- Kim, S. N., Cavedon, L., and Baldwin, T. (2010). “Classifying Dialogue Acts in One-on-one Live Chats.” In *Proceedings of EMNLP*, pp. 862–871.
- Kim, Y. (2014). “Convolutional Neural Networks for Sentence Classification.” In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1746–1751.
- 国立国語研究所 (2001). 名大会話コーパス. 科学研究費基盤研究 (B)(2) 「日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究」(平成 13 年度～15 年度), <http://pj.ninjal.ac.jp/conversation/nuc.html>.
- Libin, A. V. and Libin, E. V. (2004). “Person-robot interactions from the robopsychologists’ point of view: the robotic psychology and robotherapy approach.” In *Proceedings of the IEEE*, **92** (11), pp. 1789–1803.
- 前田英作, 南泰浩, 堂坂浩二 (2011). 人口ロボット共生におけるコミュニケーション戦略の生成. 日本ロボット学会誌, **29** (10), pp. 887–890.
- 目黒豊美, 東中竜一郎, 杉山弘晃, 南泰浩 (2013). 意味属性パターンを用いたマイクロブログ中の発言に対する自動対話行為付与. 研究報告音声言語情報処理 (SLP), **2013** (1), pp. 1–6.
- Meguro, T., Minami, Y., Higashinaka, R., and Dohsaka, K. (2014). “Learning to Control Listening-oriented Dialogue Using Partially Observable Markov Decision Processes.” *ACM Transactions on Speech and Language Processing*, **10** (4), pp. 1–20.
- Milajevs, D. and Purver, M. (2014). “Investigating the Contribution of Distributional Semantic Information for Dialogue Act Classification.” In *Proceedings of the 2nd Workshop on*

*Continuous Vector Space Models and their Compositionality (CVSC)*, pp. 40–47.

南泰浩, 東中竜一郎, 堂坂浩二, 目黒豊美, 森啓, 前田英作 (2012). 対話行為タイプ列 Trigram による行動予測確率に基づく POMDP 対話制御. 電子情報通信学会論文誌. A, 基礎・境界, **95** (1), pp. 2–15.

西田公昭 (1992). 対話者の会話行為が会話方略ならびに対人認知に及ぼす効果. *The Japanese Journal of Psychology*, **63** (5), pp. 319–325.

関野嵩浩, 井上雅史 (2010). 発話に対する拡張談話タグ付与. 第 6 回情報処理学会東北支部研究会報告.

Sugiyama, H., Meguro, T., Higashinaka, R., and Minami, Y. (2013). “Open-domain Utterance Generation for Conversational Dialogue Systems using Web-scale Dependency Structures.” In *Proceedings of SIGDIAL*, pp. 334–338.

Takahashi, K., Takamura, H., and Okumura, M. (2007). “Estimation of Class Membership Probabilities in the Document Classification.” In *Proceedings of the 11th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD)*, pp. 284–295.

## 付録

### A 対応表

正解の対話行為と提案手法のうち最も F 値の高い  $Pro_b$  が選択した対話行為の対応表を表 15 に示す.

表 15 正解の対話行為と  $Pro_b$  の出力との対応表

		(Pro <sub>b</sub> の出力)								
		$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$	$d_7$	$d_8$	$d_9$
(正解)	$d_1$	4,622	22	24	2	94	65	18	15	10
	$d_2$	93	466	28	0	2	1	1	28	0
	$d_3$	55	37	252	0	5	3	2	3	0
	$d_4$	5	0	0	184	9	8	3	0	0
	$d_5$	106	5	1	6	655	2	4	2	0
	$d_6$	149	1	2	13	6	649	100	9	0
	$d_7$	73	1	6	5	9	79	203	0	0
	$d_8$	213	79	3	0	17	15	1	128	0
	$d_9$	67	2	0	0	11	0	0	0	18

$d_1$ : 自己開示,  $d_2$ : 質問 (YesNo),  $d_3$ : 質問 (What),  $d_4$ : 応答 (YesNo),  $d_5$ : 応答 (平叙),  $d_6$ : あいづち,  $d_7$ : フィラー,  $d_8$ : 確認,  $d_9$ : 要求

## 略歴

**福岡 知隆**：2010年東京工科大学コンピュータサイエンス学科コンピュータサイエンス学部卒業。2012年同大学院バイオ・情報メディア研究科修士課程修了。2017年北陸先端科学技術大学院大学情報科学研究科博士課程修了。同年株式会社 Nextremer AI エンジニア。現在に至る。博士（情報科学）。電子情報通信学会会員。

**白井 清昭**：1993年東京工業大学工学部情報工学科卒業。1998年同大学院情報理工学研究科博士課程修了。同年同大学院助手。2001年北陸先端科学技術大学院大学情報科学研究科助教授。現在に至る。博士（工学）。統計的自然言語解析に関する研究に従事。情報処理学会，人工知能学会，電子情報通信学会各会員。

(2016年9月5日 受付)

(2016年12月6日 再受付)

(2017年4月22日 再々受付)

(2017年6月14日 採録)