個々の対話行為の特徴を考慮した自由対話における対話行為推定

福岡 知隆　　白井 清昭
北陸先端科学技術大学院大学 情報科学研究科
{s1320010, kshirai}@jaist.ac.jp

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．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いくつも提案されている．しかし，機械学習に用いる特徴量を設定する際，個々の対話行為の特徴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量のセットを１つ設定する．しかし，特徴量の中には，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量は重要だが，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量とは考え難い．

本論文では，上記の問題に対し，対話行為毎に適切な特徴量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

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

単語 uni-gram は語順を考慮していないため，Milajevs らは単語 bi-gram を特徴量として用い，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 [7]．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量として利用し，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量として，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 [2]．他の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量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，特徴量として発話文字数，内容語数，発話順番を提案し，磯村の手法 [2] の特徴量にこれらを１つ以上追加したモデルを評価した [8]．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特徴と，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量として提案した [4]．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．

これらの先行研究では，機械学習のために用いる特徴量のセットは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，どの特徴量がどの対話行為の推定に有効に働くかなど，特徴量と対話行為の関係については議論されていない．

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．後述するように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，学習のための特徴量のセットはあらかじめ一律に定められている．しかし，全ての特徴量が全ての対話行為の分類に必要というわけではなく，ある特徴量が特定の対話行為の分類に貢献しないことがある．そのような特徴量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量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否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

徴量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．分類器によって該当すると判定された対話行為のうち，信頼度の最も高い対話行為を選択し，それを最終の出力とする．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採用し，学習ツールとして LIBLINEAR[1] を用いた．判定の信頼度は LIBLINEAR が出力する確率を用いた．

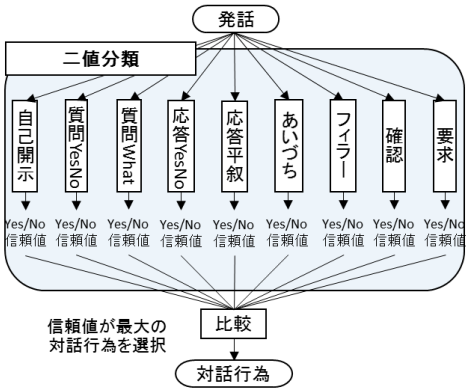

図 1: 提案手法の流れ

3.1 対話行為の定義

対話行為の定義としては SWBD-DAMSL[3] 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 [5] も提案されてはいるが，本研究では独自に定義した 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.

表 1: 対話行為の定義

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

3.2 特徴量

対話行為の言語的特徴を考慮し，表 2 に示す 28 個の特徴量を提案する．単語 n-gram を用いた特徴量 (f_1, f_2, f_8, f_9, f_{13}) では $n = 1, 2, 3$ とする.「前発話」とは直前の相手の発話を指す．f_{13}, f_{14} 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．f_{15}, f_{16}, f_{17}, f_{18}, f_{19}, f_{24} 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_{21}, f_{22}, f_{23} における「自立語繰り返し」とは，相手の前発話の自立語が現在の発話で繰り返し用いられるかを表わす．繰り返される自立語が文末に出現する場合を f_{22}，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_{23} として区別している．f_{25}, f_{26} 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．f_{27} 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．f_{28} 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.

また，詳細は省略するが，一部の特徴量については 2 つの特徴量を組み合わせたものも学習に用いた.

3.3 最適な特徴量の選択

個々の対話行為に対し，表 2 に示した特徴量の中から，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量のセットを決める．特徴量の有効性の判定は，全特徴量を用いた場合と，1 つの特徴量を除いた場合を比較し，開発データにおける後者の対話行為推定の F 値が低ければ，その特徴量は有効であるとみなす．有効でない特徴量は削除する．これを全ての特徴量について行い，1 つ以上の特徴量が削除されたら，残された特徴量を新たに全特徴量の集合とみなして同様の操作を行う．これを特徴量が削除されなくなるまで繰り返す.

3.4 過去の対話行為列の長さの最適化

特徴量 f_6 と f_7 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．これまで，過去の発話の対話行為を特徴量とした先行研究はあったが，我々の知る限り，対話行為列の長さを対話行為毎に最適化した研究例は報告されていない.

本研究では，相手もしくは話者自身の過去の N_h 個の発話の対話行為の列を特徴量と定義し，N_h を対話行為毎に最適化する．また，N_h の値が大きいときには素性数が増えるため，素性選択を行う．具体的には，素性と対話行為の分類クラスの相関の強さを χ^2 値で測り，それが閾値 T_h よりも小さい素性を削除する．ここでは，$N_h = 1, 2, 3, 4, 5$ ならびに $T_h = 0, 1, 5, 10$ とし，開発データでの F 値が最大となる N_h と T_h を選択する.

表 2: 対話行為推定のための特徴量

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

表 3: 選択された特徴量

特徴量	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}
自己開示					✓	✓	✓	✓		✓						✓		✓		✓			✓		✓	✓	✓	
質問 (YesNo)	✓		✓			✓		✓	✓		✓		✓	✓		✓			✓	✓			✓	✓				✓
質問 (What)	✓			✓		✓	✓	✓			✓			✓		✓			✓		✓							
応答 (YesNo)					✓	✓		✓		✓					✓	✓	✓	✓					✓	✓	✓			✓
応答 (平叙)					✓	✓			✓				✓	✓	✓	✓		✓	✓	✓	✓			✓		✓		✓
あいづち	✓		✓		✓	✓			✓		✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
フィラー	✓				✓	✓	✓		✓	✓						✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
確認	✓			✓		✓	✓	✓	✓				✓		✓	✓	✓				✓		✓	✓			✓	
要求	✓				✓		✓	✓		✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

4 評価実験

4.1 コーパス

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対話コーパス [6] を用いる．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．コーパスをおよそ 80%，10%，10%に分割し，それぞれ訓練データ，開発データ，テストデータとした．開発データは最適な特徴量の選択やパラメタの最適化のために用いる．

4.2 特徴量の選択

対話行為毎に選択された特徴量の一覧を表 3 に示す．なお，ここでは特徴量 f_6,f_7 のパラメタ N_h と N_t の最適化はしておらず，$N_h = 1$，$T_h = 0$ としている．対話行為毎に最適な特徴量が異なることがわかる．特に，f_{16}(質問 (What) キーワード)，f_6(相手の過去の発話の対話行為) などがほとんどの対話行為で選択されており，重要な特徴量であることがわかった．一方，f_2(前発話の単語 n-gram) は全ての対話行為で選択されていない．f_4(前発話の自立語) も多くの対話行為で選択されていないことから，前の発話の内容は対話行為推定のための有効な手がかりではないと言える．

4.3 パラメタの最適化

特徴量 f_6 と f_7 について，3.4 項で述べたように N_h と T_h の最適化を行った．結果を表 4 に示す．f_6(相手の過去の発話の対話行為) については，2 つ以上前からの発話の対話行為の列を特徴量とした方が結果がよく，1 つ前の発話の対話行為のみを用いたとき ($N_h = 1$) と比べて F 値が 2%程度向上することがわかった．これは，対話行為列によって対話の流れのような情報をモデルに反映することができるためと考えられる．f_7(話者の過去の発話の対話行為) についても，$N_h = 1$ となる対話行為もあるが，より長い履歴を用いた方がよい対話行為の方が多い．

表 4: 過去の発話の対話行為の特徴量のパラメタ

対話行為	f_6(相手) N_h	f_6(相手) T_h	f_7(話者) N_h	f_7(話者) T_h
自己開示	2	1	3	0
質問 (YesNo)	4	0	1	0
質問 (What)	4	0	4	5
応答 (YesNo)	2	0	1	0
応答 (平叙)	2	10	1	0
あいづち	5	1	3	0
フィラー	4	0	3	0
確認	4	0	2	5
要求	2	0	2	1

4.4 対話行為推定の評価

それぞれの対話行為について，発話がその対話行為に該当するかの二値分類を行い，その精度 (P)，再現率 (R)，F 値 (F) を調べた．結果を表 5 に示す.「全て」は表 2 に示した全特徴量を使ったとき,「最適化」は表 3 の通りに選択した特徴量を用いたときの結果である．特徴量の選択は開発データを用いて行ったが，テストデータにおいても「あいづち」を除く全ての対話行為において，特徴量の最適化によって F 値が向上することを確

表 5: 対話行為毎の二値分類の結果

	全て			最適化		
	P	R	F	P	R	F
自己開示	.848	.906	.858	.856	.925	.889
質問 (YesNo)	.757	.658	.704	.763	.680	.719
質問 (What)	.790	.644	.710	.787	.672	.725
応答 (YesNo)	.907	.842	.873	.872	.885	.879
応答 (平叙)	.815	.723	.767	.804	.798	.801
あいづち	.769	.721	.744	.763	.717	.739
フィラー	.651	.298	.409	.612	.356	.450
確認	.653	.241	.351	.680	.254	.370
要求	.826	.194	.314	.714	.204	.317
マクロ平均	.780	.581	.637	.761	.610	.654

表 6: 対話行為の推定結果

	ベースライン			提案手法		
	P	R	F	P	R	F
自己開示	.844	.945	.892	.854	.953	.901
質問 (YesNo)	.753	.748	.750	.761	.746	.754
質問 (What)	.758	.692	.723	.800	.683	.737
応答 (YesNo)	.921	.842	.880	.856	.852	.854
応答 (平叙)	.798	.717	.755	.798	.827	.812
あいづち	.731	.758	.744	.768	.720	.743
フィラー	.599	.346	.438	.621	.410	.494
確認	.624	.278	.385	.704	.289	.410
要求	.792	.194	.311	.652	.153	.248
マクロ平均	.758	.613	.653	.757	.626	.661
マイクロ平均	.808	.808	.808	.821	.821	.821

認した.

次に，図 1 に示した提案手法で対話行為を推定した結果を表 6 に示す．表中の「ベースライン」は，全特徴量を用い，LIBLINEAR で対話行為を分類した結果である．提案手法はベースラインと比べて，F 値のマクロ平均で 0.8%，マイクロ平均[1]で 1.3%の改善が見られた．また，マクネマー検定の結果，提案手法とベースラインは有意水準 1%で有意差があることがわかった．

対話行為毎に提案手法とベースラインを比較すると，全ての対話行為について F 値が改善されているわけではない．なかでも「要求」の F 値は 6.3%低下している．「要求」の二値分類の F 値は 31.7%(表 5) であるが，それが表 6 では 24.8%に大きく低下しているのは,「要求」の分類器が出力する判定の信頼度が他の対話行為に比べて低いためである．すなわち，対話行為が「要求」である発話に対し,「要求」の二値分類器は正しく判定できているが，他の対話行為に対する判定の信頼度の方が高いため,「要求」以外の対話行為が誤って選択されている．

5 おわりに

本論文は，自由対話における発話の対話行為を自動推定するために，対話行為毎に最適な特徴量を選択することで F 値を向上させる手法を提案した．評価実験の結果，全ての対話行為に対して同一の特徴量を使用するときと比べて，対話行為推定の F 値を 1.3%向上させることができた．また，対話行為毎に最適な特徴量のセットが異なること，過去の発話の対話行為列を特徴量とするときはその長さを調整することが重要であるとの知見を得た．提案手法には，対話行為毎に判定の信頼度が大きく異なるため，判定の信頼度が全般的に低い対話行為が選択されにくいという問題点がある．判定の信頼度を比較する際，対話行為毎に重みを与える方法などを検討し，対話行為推定の F 値を更に向上させることが今後の課題となる．

[1] この場合のマイクロ平均は，精度，再現率，F 値のいずれも正解率 (システムが選択した対話行為が正解と一致する割合) と一致する．

参考文献

[1] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, Vol. 9, pp. 1871–1874, 2008.

[2] 磯村直樹, 鳥海不二夫. 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, Vol. 92, No. 11, pp. 795–805, 2009.

[3] Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. Technical report, Institute of Cognitive Science Technical Report, 1997.

[4] Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. Classifying dialogue acts in one-on-one live chats. In *Proceedings of EMNLP*, pp. 862–871, 2010.

[5] Toyomi Meguro, Yasuhiro Minami, Ryuichiro Higashinaka, and Kohji Dohsaka. Learning to control listening-oriented dialogue using partially observable markov decision processes. *ACM Transactions on Speech and Language Processing*, Vol. 10, No. 4, pp. 1–20, 2014.

[6] 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).

[7] Dmitrijs Milajevs and Matthew Purver. 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, 2014.

[8] 関野嵩浩, 井上雅史, 横山晶一. 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, 2010.