対話行為の分布を利用した雑談対話システムの評価指標

榮田亮真　井手竜也　村田栄樹　河原大輔
早稲田大学理工学術院
{s.ryoma6317@akane.,t-ide@toki.,eiki.1650-2951@toki.,dkw@}waseda.jp

概要

本論文では，雑談対話システムのふるまい全体を考慮して評価を行うための新たな指標として，発話が持つ役割である対話行為の分布を利用することを提案する．人間とシステムの各発話に対話行為が付与された対話コーパスの分析から，人間は多様な対話行為を使い分けていること，また，対話相手の発話を踏まえて対話行為を決定していることがわかった．分析に基づき，対話行為のエントロピーと，発話の対話行為が，対応する応答の対話行為の決定に与える相互情報量で雑談対話システムを評価する．提案指標を様々な雑談対話システムに適用した評価実験によって，指標の妥当性が確認された．

1　はじめに

現在の雑談対話システムの評価方法は課題が残ったまま利用されている．評価は人手評価と自動評価に分けることができ，人手評価はクラウドソーシングによるもの，自動評価は BLEU [1]，Distinct [2] などの指標によるものが一般的である．さらに近年では，BERT [3] などの事前学習モデルを人手評価データで Fine-tuning したモデルが提案されている [4, 5]．

既存の評価指標の多くは，雑談対話システムの 1 ターンや 1 対話の応答を評価するものである．このような評価指標はシステムのふるまい全体を見ていないため，決まった数パターンの応答しか返すことができないシステムを高く評価することがある．また，Distinct はふるまい全体を評価する指標だが，表面的な語彙の多様性をはかるだけである．

本論文では，既存手法の問題点に鑑み，雑談対話システムのふるまい全体を見て，Distinct よりも適切に，システムを評価することのできる評価指標を提案する．新たな評価指標は，対話における各発話の役割を表す対話行為 [6] の分布に注目したものである．指標の考案のためまず，人間とシステムの発話に対話行為が付与された対話コーパスを構築し，分析を行う．コーパスの分析により，システムと比べて人間は対話行為を多様に使い分けていること，また，対話相手の発言を踏まえて，対話行為を決定していることがわかった．分析に基づき，対話行為の多様性を表すエントロピー (DAE)[1] と，どれだけ相手の発言の対話行為を考慮しているかを表す相互情報量 (DAMI)[2] に注目することで，システムを評価することを提案する．DAE は Distinct とは異なり，おうむ返しばかりをするシステムに低いスコアを与える．DAMI は応答だけでなく，その直前の発話にも注目したもので，対話の評価として Distinct よりも適切であると考えられる．DAE，DAMI の算出には発話に対話行為を付与することが必要だが，付与方法として，人手付与と対話行為分類モデルによる自動付与の両方を検討する．

提案指標の妥当性検証のため，様々な雑談対話システムに対して，提案指標を算出する実験を行った．実験から，提案指標の DAE，DAMI が対話システムの性能を反映した指標であることがわかった．

2　関連研究

既存の雑談対話システムの評価方法の多くが，1 つのユーザ発話とそれに対応するシステム応答を見て評価を行うターンレベルの評価である．ターンレベルの関連度の人手評価 [7] は，おうむ返しやありきたりでつまらない応答についても高い評価を与える．これは，おうむ返しやありきたりな応答を多く返すことで知られている深層学習による雑談対話システム [8] の評価指標として不適切である．

対話行為（Dialog Act）とは，対話において各発話が持つ役割である．Stolcke ら [6] は 42 種類の対話行為を提案し，対話行為を付与した対話コーパスである SwDA を構築した．対話行為が付与された他のコーパスとしては，日常的な対話を収集した DailyDialog [9] などが存在する．また，深層学習モ

1） Dialog Act Entropy
2） Dialog Act Mutual Information

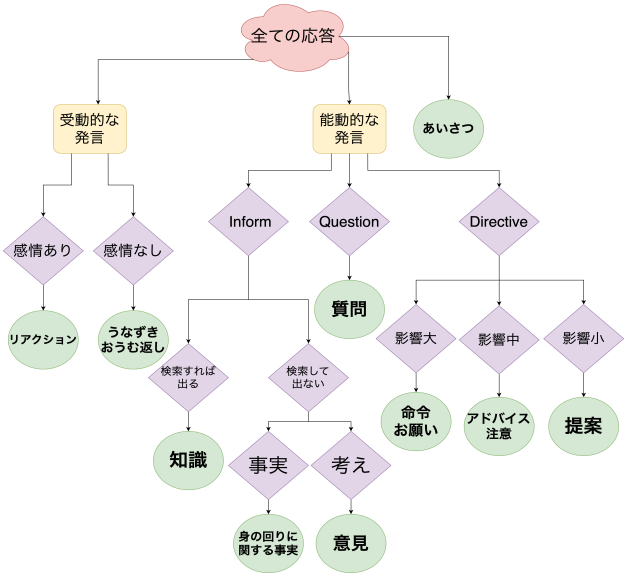


図 1: 対話行為の体系（Inform、Question、Directive は DailyDialog で提案されたもの）

表 1: 対話行為の分布（太字は各対話行為について人間とシステムのどちらが多いかを示している）

対話行為	人間発話	システム発話
あいさつ	366	**737**
うなずき・おうむ返し	446	**827**
リアクション	**1,080**	646
知識	**115**	83
身の回りに関する事実	**790**	595
意見	3,116	**3,245**
質問	**132**	57
命令・お願い	**162**	135
アドバイス・注意	199	**252**
提案	**223**	119
その他	**79**	33

デルによって対話行為を分類することを目指した研究も存在する [10, 11, 12].

3 対話行為データセットの構築

雑談対話システムの評価指標の考案のため，人間とシステムの各発話に対話行為が付与された対話コーパスを構築する. このコーパスを対話行為データセットと呼ぶ.

3.1 対話行為の体系

本論文では，10 種類の対話行為で構成される体系を提案する (図 1)．能動的な発言は、DailyDialog [9] で提案された対話行為をもとに分類した.

3.2 応答生成モデル

対話システムによる発話の対話行為を収集するために，T5 [13] の日本語版[3)] を Fine-tuning したシステムを用いてシステム発話を得る. Fine-tuning に用いるデータセットは Twitter から Twitter API を用いて収集したツイートリプライ対で，800,000 ペアである.

3） Hugging Face Hub: sonoisa/t5-base-japanese

3.3 対話行為のアノテーション

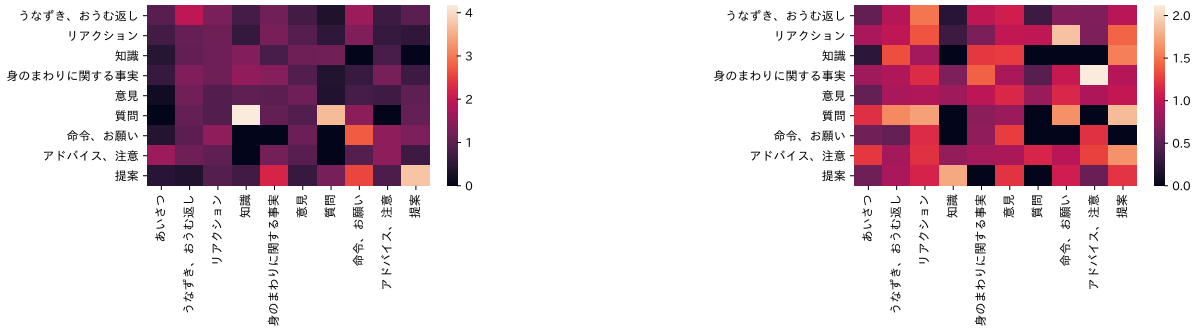
3.1 節に示した対話行為の体系に基づき，クラウドソーシングでアノテーションを行う. アノテーションの対象は，ある発話に対する人間の応答とシステムの応答それぞれ 6,820 文である．発話と人間による応答は Twitter から TwitterAPI を用いて収集したマルチターン対話で，システムによる応答は 3.2 節の応答生成モデルから得たものである．対話をクラウドワーカに見せ，最後の発話に関して，当てはまる対話行為を全て選択してもらう. 1 つの発話につき 5 人に尋ね，2 票以上かつ最多の票を集めた対話行為を採用する. 複数の対話行為が採用対象になった場合，それらの中からランダムに 1 つ選択する. 収集したデータセットの分布を表 1 に示す. 人間の方が多様に対話行為を使い分けていることがわかる.

3.4 対話行為の遷移

対話においては，ある対話行為の後には特定の対話行為が現れやすいといった傾向が存在すると考えられる．その傾向における人間とシステムの違いを見るため，構築したデータセットにおいて対話行為の遷移を分析する．遷移を示したヒートマップを図 2 に示す．ヒートマップの各行は，その行名の対話行為の後にどの対話行為が来るかの条件つき確率分布を，事前確率分布で補正した値 $\frac{P(a_{t+1}|a_t)}{P(a_{t+1})}$ を表している．a_t は対話における t 番目の発話の対話行為を表す．事前確率分布による補正をすることで，それ以外の状況と比較したときに，特定の状況下で，ある対話行為が現れやすいことを表現した値となる．ヒートマップから，人間はシステムに比べて，対話相手の対話行為を踏まえてどんな対話行為を返すべきかを決定しているという傾向がわかる.

4 対話行為分布を利用した雑談対話システムの評価

3 節の対話行為データセットの分析を踏まえて，雑談対話システムを評価するための 2 つの指標を提案する.



(a) 人間の発話の対話行為遷移

(b) システムの発話の対話行為遷移

図 2: 対話行為遷移を示したヒートマップ

表 2: 人手付与された対話行為に基づく評価指標による評価結果と Distinct（太字はそれぞれの比較対象についてスコアが最大のものを示している）

システム	DAE	DAMI	Distinct
T5-8k	2.24	0.0709	0.0377
T5-80k	2.27	0.0703	0.0855
T5-400k	2.34	**0.107**	0.0819
T5-800k	**2.41**	0.104	**0.0961**
mT5-small	1.96	0.063	0.0446
mT5-base	**2.15**	**0.076**	0.0539
おうむ返し	0	0	0.151
Human	2.48	0.237	0.160

4.1 エントロピー (DAE)

3.3 節による分析から，人間は対話の中で，さまざまな対話行為を使い分けていることがわかる．そのため，対話システムの評価指標としてシステム発話の対話行為の多様性が利用できると考えられる．分布の多様性は，エントロピー[4)]で計算される．

4.2 相互情報量 (DAMI)

3.4 節の分析から人間は，相手の発言の内容に応じて対話行為を使い分けていることがわかる．このことは，直前の発話の対話行為が対象発話の対話行為の決定に与える相互情報量[5)]を用いて定量的に表現することができる．

5 評価指標の検証

対話行為を付与する方法として人手付与と対話行為分類モデルによる自動付与の 2 つがあり，それぞれにおける評価指標の妥当性を検証する．

5.1 人手付与による評価

5.1.1 実験設定

本実験では，対話行為を 3.3 節と同様に，クラウドソーシングで収集する．評価指標算出のためのシステム発話は 6,000 文であり，3.3 節でシステム発話を得た Twitter 上のマルチターン対話 6,820 文のうち，6,000 文の応答として得たものである．対話システムの構築は事前学習モデルである T5[6)] [13]，mT5[7)] [14] を Fine-tuning して行う．Fine-tuning に用いるデータセットは，3.2 節で用いたものと同じツイートリプライ対である．日本語 T5 は単一のモデルサイズしか公開されていないため，学習データ量をシステムの性能とみなし，提案指標の比較を行う．データセット全量を用いたシステムに加えて，8,000 ペア, 80,000 ペア, 400,000 ペアを使用して Fine-tuning したシステムも構築する．なお，データセット全量を用いたシステムとはすなわち，3.2 節のシステムである．mT5 については，モデルパラメータの多さをシステムの性能とみなす．また，入力をそのまま返すおうむ返しシステムと人間による発話についても提案指標を算出する．

5.1.2 実験結果

実験結果を表 2 に示す．T5 について，Fine-tuning に用いたデータ量が多いシステムの方が高いスコアを得た．また，mT5 について，パラメータ数が多い base モデルの方が高いスコアを得た．これらの結果から，提案指標は妥当なものであると言える．さらに，人間はどちらの指標においても最も高いスコア

4) 定義は付録に示す.
5) 定義は付録に示す.
6) Hugging Face Hub: sonoisa/t5-base-japanese
7) Hugging Face Hub: google/mt5-small, mt5-base

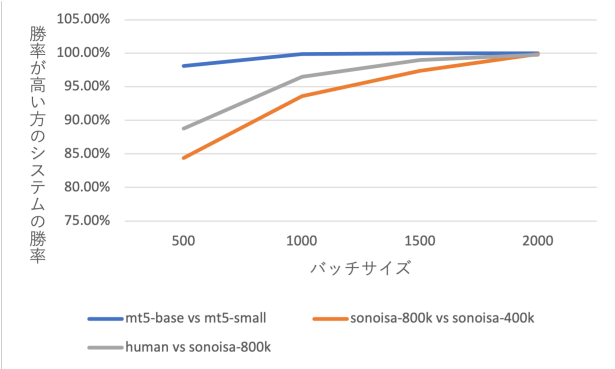


図 3: バッチサイズに対する比較結果の変化

表 3: 自動付与された対話行為に基づく対話システムの評価結果

システム	DAE
T5-8k	2.46
T5-80k	2.72
T5-400k	2.71
T5-800k	2.77
mT5-small	2.58
mT5-base	2.45
Human	2.89

を得た．Distinct はおうむ返しだけを返すシステムに対しても高いスコアを与えたが，提案指標の値は 0 である．

評価に一貫性はあるか

評価指標においては，異なるテストコーパスを用いても一貫した結果であることが重要であるため，検証を行う．比較したい 2 つのシステムについて，クラウドソーシングで収集した 6,000 発話からバッチサイズ分の発話をサンプリングし，エントロピーを計算する．サンプリングを繰り返し，エントロピーの大小が一貫しているかを確認する．サンプリング時には，同一の発話に対応する 2 つのシステムの発話をサンプリングすることに注意する．比較は以下の 3 ペアに関して行う．

- mT5-base vs mT5-small
- T5-800k vs T5-400k
- human vs T5-800k

バッチサイズを 500, 1000, 1500, 2000 と変更させる．サンプリングの繰り返し回数は 20,000 回とする．実験結果を図 3 に示す．2,000 文のテストコーパスを用いれば，一貫した結果が得られることがわかる．

5.2 自動付与による評価

発話の対話行為の分類モデルを構築し，5.1 節と同様の実験を行う．

5.2.1 実験設定

モデルの構築は事前学習モデル BERT[8] を Fine-tuning して行う．学習コーパスは対話行為データセットで，人間の発話とシステムの発話の両方を用いる．表 1 から，学習データ数は対話行為ごとに偏りがある．データ数の偏りは学習に悪影響であると考えられるため，各対話行為のデータ数を，最多の対話行為である「意見」に合わせるアップサンプリングを行う．学習用と検証用にデータセットを 4:1 に分割し，検証用のデータセットに対してクロスエントロピーロスが最小であったエポックを採用する．テストコーパスは Twitter 上のマルチターン対話を各対話システムに入力して得た応答で，学習に用いたデータセットとは異なる 3,880 文である．

5.2.2 実験結果

自動付与した対話行為を利用して評価指標を計算した結果を表 3 に示す．T5 と人間については妥当な結果が得られているのに対し，mT5 は正しく評価できなかったことがわかる．これは学習データの発話が T5 と人間によるものであることが原因と考えられる．自動付与による評価を実用化するためには，分類モデルの改善が必要である．

6 おわりに

本論文では，雑談対話システムのふるまい全体を評価することのできる指標を提案した．提案指標は，システム発話の対話行為の分布から算出されるエントロピーと相互情報量であり，データセットの分析によって得た人間とシステムの発話の特徴に基づくものである．様々な対話システムを対象に提案指標を算出する実験によって，提案指標がシステムの良さを反映した妥当な指標であることがわかった．提案指標は応答の対話行為だけを見ており，内容の質を見ていない．応答がその対話行為の応答として質が高いかを評価する手法を確立し，統合することが今後の課題である．

8) HuggingFace Hub: cl-tohoku/bert-base-japanese-whole-word-masking

謝辞

本研究は 2022 年度国立情報学研究所 CRIS 委託研究の助成を受けて実施した.

参考文献

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[2] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 110–119, San Diego, California, June 2016. Association for Computational Linguistics.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[4] Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara. Designing precise and robust dialogue response evaluators. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 26–33, Online, July 2020. Association for Computational Linguistics.

[5] Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In **Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation**, pp. 82–89, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[6] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. **Computational Linguistics**, Vol. 26, No. 3, pp. 339–374, 2000.

[7] Sarah E. Finch and Jinho D. Choi. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In **Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue**, pp. 236–245, 1st virtual meeting, July 2020. Association for Computational Linguistics.

[8] Shaojie Jiang and Maarten de Rijke. Why are sequence-to-sequence models so dull? understanding the low-diversity problem of chatbots. In **Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI**, pp. 81–86, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[9] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In **Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 986–995, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[10] Ji Young Lee and Franck Dernoncourt. Sequential short-text classification with recurrent and convolutional neural networks. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 515–520, San Diego, California, June 2016. Association for Computational Linguistics.

[11] Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers**, pp. 2012–2021, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[12] Vipul Raheja and Joel Tetreault. Dialogue Act Classification with Context-Aware Self-Attention. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 3727–3733, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[13] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[14] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 483–498, Online, June 2021. Association for Computational Linguistics.

表 4: 意見の分布

意見の対象	属性	人間発話		システム発話	
対話相手	ポジティブ	1,127	38%	1,467	**47%**
	ネガティブ	320	11%	391	**13%**
	ニュートラル	792	**27%**	648	21%
それ以外	ポジティブ	308	10%	311	10%
	ネガティブ	237	**8%**	151	5%
	ニュートラル	181	**6%**	129	4%

A 意見の細分化

表 1 から，人間とシステムどちらにおいても，「意見」の発話が多いことがわかる. そこで，「意見」を細分化して，さらなる特徴の違いを見る.「意見」を，「意見の対象」と，「意見の属性」によって 6 種類に分類する. 分類の詳細を表 4 に示す. アノテーションの対象は 3.3 節の収集で「意見」の対話行為が付与されたものである. マルチターンの対話をクラウドワーカに見せ，最後の発話に関して，6 種類の選択肢から 1 つを選択してもらう. 1 つの発話に関して 5 人に尋ね，2 票以上かつ最多の票を集めた対話行為を採用する. 複数の対話行為が採用対象になった場合，ランダムに 1 つの対話行為を選択する. 収集したデータセットの分布を表 4 に示す. システムは対話相手に対するポジティブな意見を多く生成することがわかる. これはシステムからは同意の発話が多いことを示している. また，システムが対話相手以外に対する意見を生成することは少ないことがわかる. これは，対話相手以外に対する意見を生成するには，直前の発話に含まれないが対話相手との間で共有している知識や，世界知識が必要であるためと考えられる. 意見という 1 つの対話行為を詳細に見ても，人間の方が対話行為に多様性があることがわかる.

B 対話行為遷移の例

図 2a に示したヒートマップにおける，頻出の遷移の会話例を以下に示す.

(1) 質問 → 質問
a.なんか小さい頃公園で飲んだ水道みたいな味しました
b.鉄の味するってやつすか **(質問)**
c.カルキなんですかね．ネバっとする感じの **(質問)**
お互いがわからないことについて話している状況

(2) 提案 → 提案
a.音声録音して送ってくれ
b.土日飯行くならその時にでも **(提案)**
c.土曜は無理だから日曜なら **(提案)**
お互いに提案を繰り返し，約束を決めているような状況

(3) 命令・お願い → 命令・お願い
a. 久しぶりのお休みだから娘達にご飯行こうって誘ってもフラれてしまった私.. あちこち生誕準備買いにいこ
b. たまにはご馳走して **(命令・お願い)**
c.じゃあ今すぐ日本橋に来てください **(命令・お願い)**
お願いされたことについて交換条件を提示しているような状況

表 5: 対話行為分類モデルの性能

対話行為	Precision	Recall	F1
あいさつ	0.80	0.90	0.85
うなずき・おうむ返し	0.22	0.39	0.28
リアクション	0.29	0.29	0.29
知識	0.05	0.05	0.05
身の回りに関する事実	0.27	0.52	0.35
意見	0.73	0.39	0.51
質問	0.30	0.46	0.37
命令・お願い	0.26	0.42	0.32
アドバイス・注意	0.24	0.36	0.29
提案	0.19	0.29	0.23

C 評価指標の数学的定義

DAE は以下で定義される.

$$DAE = -\sum p(a_u) \log\ p(a_u) \tag{1}$$

DAMI は以下で定義される.

$$H(a_{u_{t+1}}|a_{u_t}) = -\sum p(a_{u_t}) \sum p(a_{u_{t+1}}|a_{u_t}) \log\ p(a_{u_{t+1}}|a_{u_t}) \tag{2}$$

$$DAMI = H(a_{u_{t+1}}) - H(a_{u_{t+1}}|a_{u_t}) \tag{3}$$

D 対話行為分類モデルの性能

5.2.1 節で構築した対話行為分類モデルの性能を表 5 に示す.