関連タスクの予測確率分布を用いる soft-gated BERT による対話印象分類

竹下 智章 † 上垣外 英剛 ‡ 船越 孝太郎 ‡ 高村 大也 ‡§ 奥村 学 ‡
† 東京工業大学工学院 ‡ 東京工業大学科学技術創成研究院 § 産業技術総合研究所
{takeshita@lr., kamigaito@lr., funakoshi@lr., takamura@, oku@}pi.titech.ac.jp

1 はじめに

チャットアプリや SNS における発話内容から対話相手が対話内容に抱く印象を推測することは，状況に応じた適切な応答を選択する上で有用である．**対話印象**とは，人–チャットボットの対話において，システム (チャットボット) がユーザ (人) とできるだけ長く円滑な対話を続けるための指標である．本研究における対話印象分類とは，対話文を以下の 2 クラスに分類することを目的とするタスクである．

ポジティブ印象 ユーザ発話から，ユーザが対話に対し良い印象を持っていると判断される状態

ネガティブ印象 ユーザ発話から，ユーザが対話に対し悪い印象を持っていると判断される状態

対話印象分類のように対話文に対する分類を行うタスクでは，発話内容を表した文のベクトル表現を用いる他に発話内の単語のもつ感情極性 [1] や単語同士の関係性 [2] などの特徴を利用する手法がある．しかし本研究は日本語対話文を対象にしており，このような言語的資源を利用することが難しい．

一方，対話破綻検出は対話印象分類と同じく人–チャットボット対話のある文脈において，システムの発話が矛盾している，文脈から唐突である，質問に答えてないなどの対話上の破綻を起こしていることを検出する分類タスクである [3]．対話破綻検出データセット [4] の対話文に 4.1.1 節で後述する対話印象データセットと同一の基準で対話印象のアノテーションを付した際の内訳を表 1 に示す．表 1 にあるように，対話破綻検出タスクでは，破綻なし (O)，判断し難い (T)，破綻 (X) の 3 クラス分類を行う．表 1 から，システムの発言により対話破綻が生じている対話では，ユーザがネガティブな印象を抱いていることが多い傾向にあることが分かる．

本研究において我々は，対話内容に持つ印象にとって対話破綻の影響が大きいと推測し，対話破綻検出タスクの予測結果を援用した対話印象分類手法を提案する．我々の手法では，発話内容を表した文のベクトル表現を得るにあたり BERT[5] を用いるが，関連タスクの予想確率分布を用いた soft-gated BERT を新たに提案し分類モデルに利用する．実験を通して我々は，関連タスクの学習を援用する既存の手法と比較して本手法が対話印象分類において高い分類性能を持つことを示した．

表 1 対話印象と対話破綻の関連性

対話印象ラベル	対話破綻ラベル (印象ラベル内割合)		
	破綻なし (O)	判断し難い (T)	破綻 (X)
ポジティブ印象	4174 (56.2%)	312 (4.2%)	2940 (35.6%)
ネガティブ印象	94 (25.8%)	77 (20.3%)	196 (53.9%)

2 関連研究

目的となるタスクの性能を高めるために関連するタスクの学習を援用する手法にはマルチタスク学習や転移学習が存在する．対話文の分類タスクにおいてはマルチタスク学習が用いられ，関連がある事象についての分類タスク同士でモデルを一部共有する．Mcload ら [6] はユーザ意図分類とスロット抽出とのマルチタスク学習で対話行為分類の精度を向上させた．Cerisara ら [7] は対話感情分類と対話行為分類のマルチタスク学習で相互的に性能を高めようと試みたが，2 つのタスクは事象としての関連はあるものの，これらの分類タスク同士の学習過程全体における関連性が低いためにマルチタスク学習の効果が期待されるほど高くなかったことが示唆されている．関連するタスク同士で学習を共有することで相互的に補完しあうことを目的としたこれらの手法に対し，我々が提案する手法は，関連タスク同士の学習はそれぞれ独立して行われ，一方のタスクの予測結果をもう一方のタスクの学習および予測に用いる

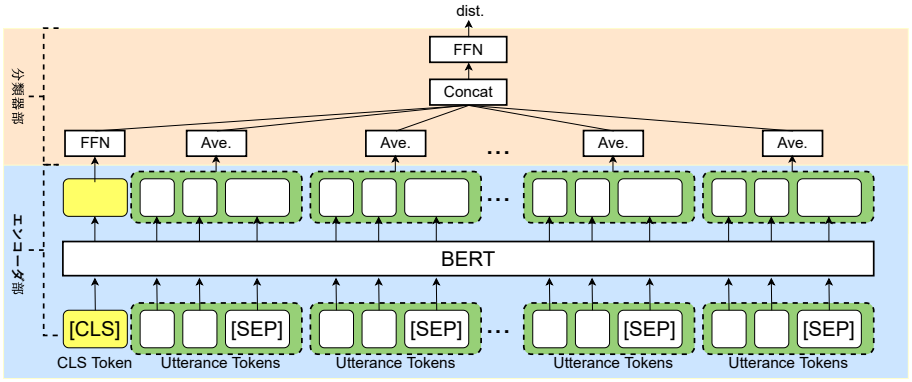

図 1　ベースラインモデルおよび対話破綻検出モデル

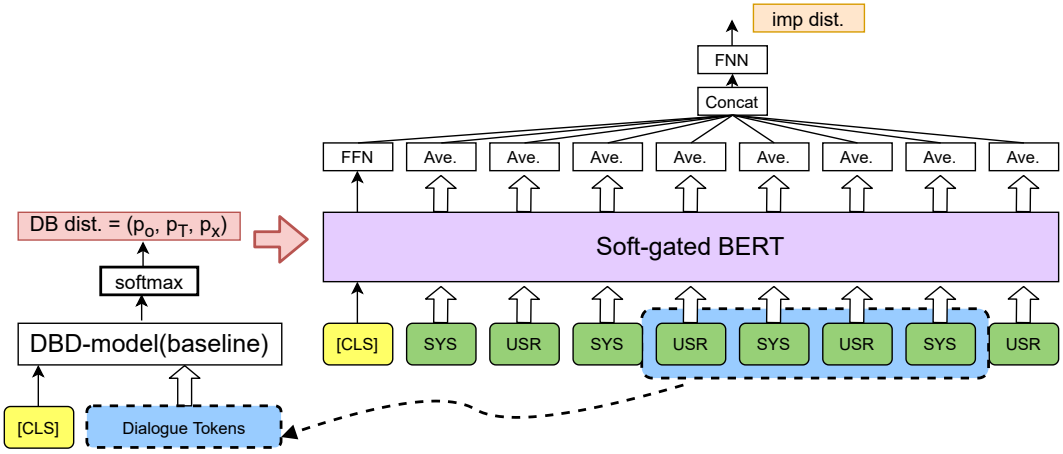

図 2　提案手法の全体図．対話破綻検出の事前学習済みモデル (DBD-model) には印象予測を行うユーザ発話の直前 4 発話が入力されて対話破綻の予測確率分布が出力される．Soft-gated BERT はこの確率分布に応じた強度でそれぞれのクラスに対応したベクトル表現を得る．Soft-gated BERT より上層のレイヤの挙動はベースラインモデルと同一である．

という相違点がある．

3　提案手法

3.1　ベースとなる手法

本研究で提案するモデルは，杉山による BERT を用いた対話破綻検出モデル [8] をベースとしている．図 1 のように，対話文内の各発話はトークン単位で分割され，発話同士は時系列順に [SEP] トークンで区切られて入力される．BERT を通して出力されたベクトルは発話単位で平均化された後結合される．結合されたベクトルは完全結合層に入力されて最終的な分類の確率分布が出力される．このモデルは対話印象分類のベースライン手法となり，4.1.2 節で後述する比較手法もこのモデルがベースとなる．

3.2　Soft-gated モデル

本研究では，ユーザが対話内容に抱く印象にとって対話破綻の影響は大きいという仮定のもと，対話破綻検出の予測結果を対話印象分類に援用する手法

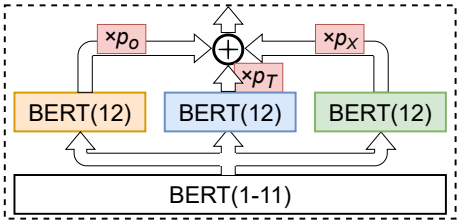

図 3　関連タスクの予測確率分布を援用した Soft-gated BERT

を提案する．この際に留意しなければならないのは，この 2 つのタスクの関連性である．これについて検討した結果，我々は，対話印象から対話破綻への一方的な依存関係があると仮定した．すなわち，対話破綻が起こっていると対話印象が悪く (=ネガティブ印象) なりやすいが，ユーザの対話印象が悪くなる際に必ずしも対話破綻が起こっているとは限らない．そのため，モデルの学習を共有することで相互的に補完し合うマルチタスク学習や転移学習のような既存の手法では過学習が起こる懸念がある．

そこで本研究では，ベースライン手法で用いられ

る BERT を図 3 のようなモデルに置き換えることで，対話破綻検出の予測確率分布に応じたベクトル表現を得るモデルを提案する．このモデルは事前学習済の BERT の最上層 (12 層目) を複製し，3 個のレイヤ (各レイヤは対話破綻検出の各クラスに対応) を用意する．11 層目から出力されたベクトルは分岐したそれぞれのレイヤに入力され，12 層目の出力はそれぞれ対話破綻の予測確率分布 (p_O, p_T, p_X) 倍の強度で足し合わされる．対話破綻検出の確率分布の予測は事前に対話破綻検出を学習したモデル (ベースラインモデルと同じ物) によって行われる (図 2)．このような機構を導入するすることで，対話破綻検出のクラスに対応する分岐レイヤには逆伝播する際の強度の違いによって学習に差異が発生することになる．通常 BERT 内の multi-head attention がそれぞれ異なる表現を自動的に学習することで柔軟な表現に対応しているが，本手法では先験的に関連タスクに対応したレイヤを用意することで，BERT の学習済み表現を大きく壊すことなく，対話の破綻状態に応じた適応学習が行われることが期待される．

4 実験

4.1 実験設定

実験に用いたデータセットおよび比較手法，評価方法について説明する．

4.1.1 データセット

対話印象分類データセット 本研究で扱う対話印象は，株式会社ホンダ・リサーチ・インスティチュート・ジャパンによる人–チャットボット対話文コーパスにアノテーションを付したデータセット[1)]の基準に基づく．対話文コーパスはシステム (チャットボット) とユーザ (人) の一対一の対話で構成される．ユーザの各発話には，第三者のアノテータによるユーザが対話に対し抱く印象のアノテーションラベルが付され，ポジティブ印象またはネガティブ印象で評価される．各対話には 10 種類のうちいずれか 1 つのトピックが設定され，システムからそのトピックに対応する話題をユーザに提示[2)]することで，特定のトピックに関する対話を行う．

表 2 対話印象分類データセットの内訳

クラス	事例数
ポジティブ印象	12695
ネガティブ印象	1861

表 3 対話破綻検出データセットの内訳

クラス	事例数
破綻なし (O)	6921
判断し難い (T)	3544
破綻 (X)	3495

対話破綻検出データセット 対話破綻検出チャレンジ 1[4]，2[9]，3[10]，4[11] のデータセットを用いる．対話破綻検出では，対話の状態を 3 クラス (O:破綻なし，T:(破綻しているか) 判断し難い，X:破綻) で分類する．

4.1.2 比較手法

提案手法の有効性を示すため，関連タスクの学習を利用する既存の学習手法と比較および考察を行った．

ベースライン手法: baseline 対話印象分類データセットでのみ学習を行い，対話破綻検出の学習による援用の有効性を検証する．後述する手法では便宜として当モデルを BERT を用いたエンコーダ部および分類器部に分けて呼称する．

転移学習: transfer ベースラインモデルで対話破綻検出を学習した後，エンコーダ部のみを流用したベースラインモデルで対話印象分類の学習を行う．

マルチタスク学習: multi-task ベースラインモデルのエンコーダ部は対話印象分類と対話破綻検出で共有し，分類器部はそれぞれのタスクでのみ学習される．転移学習とは異なり，バッチごとに対話印象分類タスクと対話破綻検出タスクを切り替えて並行して学習を行う．

特徴量の結合: feature-in 提案手法では，対話破綻検出モデルによる予測確率分布に応じた注目情報を得ているが，当比較手法ではこの確率分布を特徴量の 1 つとして，そのまま BERT から出力された後に平均化されたベクトル表現と結合して利用する．

4.1.3 実験方法

BERT は Wikipedia 日本語コーパスで事前学習済みの BERT モデル [12] を利用する．対話破綻検出のクラスは本来，T:中立 (判断し難い) のラベルを含んだ 3 クラス分類であるが，この中立ラベルを X:破綻

1) 付録に詳細を記す．
2) 例えば，「お土産」のトピックが設定された対話ではシステムが「旅先で職場の同僚向けの土産を購入するか」といった話題をユーザに提供する．

と同一として 2 クラス分類 (O:破綻なし，X: 破綻) とする設定でも同様の方法で実験を行った．この際，提案手法モデルの soft-gated BERT 内の分岐レイヤ数も 2 個になる．

提案手法と feature-in は，対話破綻検出モデルが独立しているため，これら 2 つの手法についてはこの部分が対話印象分類の学習時に適応学習 (fine-tuning:FT) を行うか否かの比較実験も行った．これは，関連タスク (対話破綻検出) 用に学習されたモデルがより目的のタスク (対話印象分類) に適するように調整されることを期待したためである．それぞれのモデルの学習率は [13] を参考に，印象分類モデル:破綻検出モデル = $5e-5 : 1e-6$ とした．

4.1.4 評価方法

各モデルの対話印象分類性能の評価は，対話印象データセット内のトピックごとに算出されたマクロ F1 スコアの平均値を比較することで行う．これは各トピックの対話に特有の単語への過学習を防ぐことと，クラス比に偏りのあるデータへの交叉検証を目的としており，10 個のトピックで分けられた対話文グループのうち，ある 1 つのグループをテストデータとした際，訓練データには残り 9 トピックの対話文グループが利用される．

4.2 実験結果

提案手法と比較手法について対話印象分類の精度を比較した結果を表 4 に示す．対話印象分類学習時に対話破綻検出モデルを適応学習するか否かで比較を行った結果を表 5 に示す．

ネガティブ印象クラスの分類について ベースラインと比較してマルチタスク学習では再現率 (Rec.) が上昇するが，適合率 (Pre.) が大きく減少しているために F1 スコアが低下している．これは仮定にもあった通り，必ずしも破綻しているから印象が悪いとは限らないので，注目するべき位置を過学習してしまったために適合率が減少したと考えられる．同様に特徴量の結合手法 (feature-in) では，適合率が上昇するが再現率が減少しているため F1 スコアが低くなった．

一方で提案手法では適合率を大きく下げずに再現率を上昇させているため，ベースラインより F1 スコアが上昇している．上述した過学習をしなかったためか，feature-in よりも性能が良い結果となった．また，対話破綻のクラス数は 2 クラスとした方が性能が良い結果となった．

表 4 関連タスクの学習を利用する手法との比較

	DBD	Positive Pre.	Positive Rec.	Positive F1	Negative Pre.	Negative Rec.	Negative F1	Macro Ave. Pre.	Macro Ave. Rec.	Macro Ave. F1
baseline	-	95.75	97.17	96.44	78.21	73.14	**75.24**	**86.98**	85.16	85.84
soft-gated	3	95.81	97.23	**96.51**	78.12	73.75	75.50	86.97	85.16	**86.00**
multi-task	3	**96.66**	96.85	96.74	74.82	**76.05**	74.95	85.74	**86.45**	85.85
transfer	3	95.24	97.36	96.17	78.22	70.28	73.53	86.73	83.77	84.90
feature-in	3	95.65	**97.44**	**96.52**	**79.38**	72.46	75.42	**87.52**	84.95	85.97
soft-gated	2	**96.10**	96.93	**96.51**	77.31	**75.09**	**75.94**	86.71	**86.01**	**86.23**
multi-task	2	95.69	97.17	96.42	**78.28**	71.58	74.63	**86.98**	84.38	85.53
transfer	2	95.56	96.83	96.18	76.70	72.63	73.75	86.13	84.23	84.97
feature-in	2	95.76	97.19	96.46	77.76	73.44	75.02	86.76	85.32	85.74

表 5 対話破綻検出モデル適応学習の有無による比較

	FT	DBD	Positive Pre.	Positive Rec.	Positive F1	Negative Pre.	Negative Rec.	Negative F1	Macro Ave. Pre.	Macro Ave. Rec.	Macro Ave. F1
baseline	-	-	95.75	97.17	96.44	78.21	73.14	75.24	86.98	85.16	85.84
soft-gated	なし	3	95.81	97.23	**96.51**	78.12	**73.75**	75.50	86.97	85.16	86.00
soft-gated	あり	3	**95.83**	96.99	96.40	77.07	73.66	74.95	86.45	**85.33**	85.68
feature-in	なし	3	95.65	97.44	96.52	79.38	72.46	75.42	87.52	84.95	85.97
feature-in	あり	3	95.58	**97.55**	**96.55**	**80.04**	72.13	**75.56**	**87.81**	84.84	**86.05**
soft-gated	なし	2	**96.10**	96.93	**96.51**	77.31	**75.09**	**75.94**	86.71	**86.01**	**86.23**
soft-gated	あり	2	95.61	97.26	96.42	78.12	72.21	74.66	86.86	84.73	85.54
feature-in	なし	2	95.76	97.19	96.46	77.76	73.44	75.02	86.76	85.32	85.74
feature-in	あり	2	95.72	96.92	96.30	76.78	73.32	74.33	86.25	85.12	85.32

対話破綻検出モデルに対する適応学習を行った場合は予想に反して，適応学習を行わなかったものよりも全体的にスコアが下がるという結果となった．

5 おわりに

本研究では，対話印象にとって対話破綻の影響は大きいという調査に基づいた仮定のもと，対話破綻検出を援用した対話印象分類手法を提案し，既存手法との比較を行った．転移学習やマルチタスク学習のような，関連タスクの学習を援用する既存の手法では分類性能が下がるのに対して本手法では性能が上昇するという有効性を示し，対話破綻の検出結果を特徴量として用いる手法よりも性能が良いことを示した．

また, 本研究において提案した soft-gated BERT は，事前学習済モデルの表現を大きく壊すことなく，関連タスクの予測確率を援用し目的タスクの予測を補助するという性質上，他のタスクにも適用することが可能と推察される．

謝辞

対話印象分類データセットを貸与いただいた株式会社ホンダ・リサーチ・インスティチュート・ジャパンに感謝いたします．

参考文献

[1] Justine Zhang, Jonathan P. Chang, and Cristian Danescu-Niculescu-Mizil. Conversations gone awry:detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pp. 1350–1361. Association for Computational Linguistics, 2018.

[2] Peixiang Zhong, Di Wang, and Chunyan Miao. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 165–176. Association for Computational Linguistics, 2019.

[3] 東中竜一郎, 船越孝太郎, 荒木雅弘, 塚原裕史, 小林優佳, 水上雅博. テキストチャットを用いた雑談対話コーパスの構築と対話破綻の分析. 自然言語処理, Vol. 23, No. 1, 2016.

[4] 東中竜一郎, 船越孝太郎, 小林優佳, 稲葉通将. 対話破綻検出チャレンジ. 言語・音声理解と対話処理研究会第 75 回研究会 (第 6 回対話システムシンポジウム), pp. 27–32. 人工知能学会, 2015.

[5] 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*, pp. 4171–4186. Association for Computational Linguistics, 2019.

[6] Sarah McLeod, Ivana Kruijff-Korbayova, and Bernd Kiefer. Multi-task learning of system dialogue act selection for supervised pretraining of goal-oriented dialogue policies. In *Proceedings of the SIGDial 2019 Conference*, pp. 411–417. Association for Computational Linguistics, 2019.

[7] Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T. Le. Multi-task dialog act and sentiment recognition on Mastodon. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 745–754, 2018.

[8] Hiroaki Sugiyama. Dialogue breakdown detection using BERT with traditional dialogue features. In *DBDC4 WOCHAT workshop*, 2019.

[9] 東中竜一郎, 船越孝太郎, 稲葉通将, 荒瀬由紀, 角森唯子. 対話破綻検出チャレンジ 2. 言語・音声理解と対話処理研究会第 78 回研究会 (第 7 回対話システムシンポジウム), pp. 64–69. 人工知能学会, 2016.

[10] 角森唯子, 東中竜一郎, 高橋哲朗, 稲葉通将. 対話破綻検出チャレンジ 3 における対話破綻検出の評価尺度の選定. 人工知能学会論文誌 35 巻 1 号 DSI-G, pp. 1–10. 人工知能学会, 2020.

[11] Ryuichiro Higashinaka, Luis F. D'Haro, Abu Shawar Bayan, Rafael Banchs, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and Joao Sedoc. Overview of the Dialogue Breakdown Detection Challenge 4. In *DBDC4 WOCHAT workshop*, 2019.

[12] 柴田知秀, 河原大輔, 黒橋禎夫. BERT による日本語構文解析の精度向上. 言語処理学会 第 25 回年次大会, pp. 205–208. 言語処理学会, 2019.

[13] Yang Liu and Mirella Lapata. Text Summarization with Pretrained Encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3730–3740. Association for Computational Linguistics, 2019.

[14] 中島圭祐, 駒谷和範, 中野幹生. 雑談対話システム構築フレームワーク pychat に基づく特定シチュエーション向け対話システム. 第 87 回人工知能学会言語・音声理解と対話処理研究会 (第 10 回対話システムシンポジウム). 人工知能学会, 2019.

[15] Mikio Nakano and Kazunori Komatani. A framework for building closed-domain chat dialogue systems. *Knowledge-Based Systems*, Vol. 204, p. 106212, 2020.

[16] Klaus Krippendorff. Reliability in content analysis. *Human Communicatuon Research*.

A 付録

A.1 対話印象分類データセットの詳細

本研究で扱う対話印象は，株式会社ホンダ・リサーチ・インスティチュート・ジャパンによる人–チャットボット対話文コーパスにアノテーションを付したデータセットの基準に基づく．対話文コーパスはシステム (チャットボット) とユーザ (人) の一対一の対話で構成される．本システムは，料理とレストランに関する話題を扱う雑談対話システム [14][15] である．ただし，[15] に記述したバージョンより一つ古いバージョンのシステムでのため，知識ベースやルールの数が若干異なる．

対話の参加者はクラウドソーシング会社である株式会社クラウドワークス経由で 105 名を募集した．参加者は，PC やスマートフォンなどの自分の端末の Web ブラウザから対話サーバーにアクセスし，14 日間で 8 回のセッションに参加した．参加者には，14 日間で 8 回のセッションに参加してもらい，1 日に 2 回以上は参加しないようにした．参加者の中には，8 回以上のセッションに参加した人もいた．また，各セッションで最低 25 回の発話をすることを求めた．各対話セッションには 10 種類のうちいずれか 1 つのトピックが設定され，システムからそのトピックに対応する話題をユーザに提示することで，特定のトピックに関する対話を行う．各ユーザの最初のセッションのトピックは「朝食」であった．残りのセッションでは，セッショントピックをランダムに選択したが，10 回以上セッションに参加しない限り，同じセッショントピックでのチャットは重複して行わなかった．参加者のうち真剣に対話に取り組んだ 96 名 (女性 48 名，男性 48 名，年齢 19〜50 代) の対話についてアノテーションを行った．96 人のうち 2 人が 5 回，4 人が 4 回のセッションを行い，残りの 90 人は 8 回以上のセッションを行ったこととなる．

ユーザの各発話には，第三者のアノテータによってユーザが対話に対し抱く印象のアノテーションが行われ，以下の評価基準で 5 段階 (-2:非常に悪い，-1:悪い，0:中立，1:良い，2:非常に良い) のラベルを付した．

- ユーザの発話から，ユーザが対話に対して良い印象を持っており，対話を継続する可能性が高いと推測され，同じ話題について同じように話を続けることができると判断された場合には，正値のスコアをアノテーションする．
- ユーザの発話から，ユーザが対話に対して良い印象も悪い印象も持っておらず，このままでは対話を止める可能性が高いと推測された場合には，0 をアノテーションする．
- ユーザの発言から，ユーザが対話に対して悪い印象を持っていること，対話をやめてしまう可能性が高いことが推測され，話題を変えるか，謝罪する必要があると判断された場合には，負値のスコアをアノテーションする．

本研究では，正値 (1, 2) のラベルを一律に” ポジティブ” クラスとし，負値 (-2, -1) のラベルを” ネガティブ” クラスとして利用した．また，0:中立のラベルは人手でも前述する 2 クラスとの明確な判別が困難なため今回は利用しなかった．

1 人のアノテータが全対話にアノテーションを行った．アノテーションされたスコアの信頼性を調査するために，16 のセッション (無作為に選ばれた 8 人の参加者に対して，少なくとも 25 回のユーザーターンを含む 2 つのセッションを無作為に選んだ) にもう一人のアノテータがアノテーションを行った．検証の結果，Krippendorff の α(順序尺度)[16] における 2 人のアノテータ間の一致率は 0.63 であり，中程度に高い値であった．