《中間報告会 2021/12/21》

破綻類型グループへのマルチラベル分類

初田 玲音

1 研究背景

現在研究が進められている雑談対話システム技術の 1 つとして，対話破綻検出技術がある．対話破綻検出とは，システムの対話破綻させる発話を検出する技術であり，その競技会である対話破綻検出チャレンジ (Dialogue Breakdown Detection Challenge: DBDC) も 2015 年から開催されている．対話システムに対して対話破綻検出を用いることで，システムが発話を行う前に，その発話が対話破綻を引き起こすか判定することが可能となり，対破破綻を事前に回避する事が出来るとされる．[1]

対話破綻検出では,「破綻ではない (o)，破綻 (x)，違和感あり (t)」の評価のみを行うが，これを発展させた研究として対話破綻エラー類型分類がある．対話破綻エラー類型分類は，対話破綻と推定された当該システム発話が，どの破綻エラー類型で破綻したかを分類するタスクである．この対話破綻エラー類型分類技術は，任意の対話システムに対し，対話システムの評価指標や，対話破綻する発話候補の制御に用いることが出来るため，重要な技術だと言える．しかし，破綻エラー類型のマルチラベル分類に関する研究は私の知るところ存在しない

そこで，本研究では，各破綻エラー類型が属するグループの特性を考慮することで，各グループの推定精度を改善し，それによって全体のグループ推定精度を向上させることを目的とする．

2 関連研究

類型情報に基づいた従来研究として，堀井ら [3] の研究がある．堀井らは，Project Next NLP の日本語対話タスクグループによる雑談対話の破綻原因類型化案に基づき，その類型毎に破綻識別器を作成し，それらを組み合わせる手法を提案した．破綻の原因を類型化した点は同様だが，破綻検出のタスクであるため．出力については「破綻ではない (o)，破綻 (x)，違和感あり (t)」の 3 種に留まる．そのため，どの破綻類型で破綻するかは考慮されていない．

対話破綻検出技術を利用した研究として，稲葉ら [4] の研究がある．この研究では，複数の応答候補が出力可能な対話システムに対話破綻検出器を組み込み，応答候補のリランキングを行うことで，応答の自然さを向上させた．対話破綻検出が対話システムの性能向上に有効であることが示されたが，加えて破綻原因を特定することで，リランキングを制御可能で且つ，説明可能になると期待出来る．

3 提案手法

3.1 破綻類型

本研究では，Higashinaka らが提案した，理論的なエラー分類方法とデータに基づくエラー分類を統合した統一的な対話破綻エラー類型化案に基づき，破綻したシステム発話の分類を行う．対話破綻エラー類型は，全部で 17 種類存在し，2 つの軸によって整理され，8 つのグループに分けられる．本研究では，17 種のエラー類型への分類ではなく，を提案，検討する．

エラー類型のグループ分けとその定義を表 3.1 に示す．

3.2 概要

提案するシステムの全体像を，図 1 に示す．

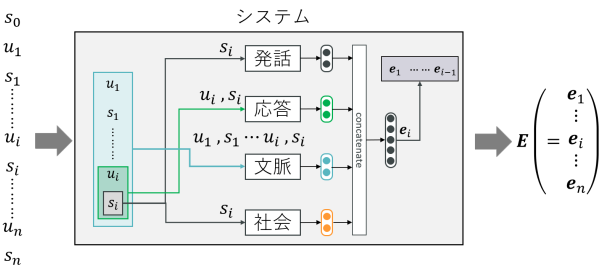

図 1: 提案モデル　全体図

表 1: エラー類型のグループ分けとその定義

	形式	内容
発話	日本語として妥当ではない発話. 文の構造を成していない発話.	単語の意味的組み合わせを誤用した発話 明らかに事実と異なる内容の発話
応答	ユーザの前向き機能を持つ発話に対応した 後ろ向き機能を有さない発話.	ユーザ発話んび対する応答の形式は正しいが 期待外れの内容の発話.
文脈	話題との関係を持つが，発話の意図が不明確な発話. 修飾句や格要素等の含むべき要素を含まない発話. 直前の話題と異なる話題へ説明なく遷移させる発話.	以前の文脈で基盤化された内容に矛盾した発話. 不要な繰り返しを行う発話
社会	言葉遣いが礼節に反し，ユーザを不快にさせる発話. 差別的な内容を含む発話.	社会通念と反するにも関わらず，根拠なく 断定或いは肯定するシステム発話.

図 1 のシステムは，ユーザ発話 u_i とシステム発話 s_i による対話全体 $C = s_0, u_1, s_1, ..., u_n, s_n$ を入力し，時刻 i 毎にエラー類型をマルチラベルに分類された横ベクトル $\mathbf{e_i}$ を縦に連結した行列 E を出力する．

提案システム内は，発話対象範囲軸を基にした 4 つのモジュールに分かれており，それぞれ入力は次のようになっている．

発話　当該システム発話のみ
応答　当該システム発話と直前のユーザ発話
文脈　当該システム発話までの対話全体
社会　当該システム発話のみ

また，それぞれのモジュールで形式，内容の 2 つのグループに分けられており，2 次元のベクトルを出力する．全てのモジュールの出力を統合し，$\mathbf{e_i}$ を各時刻 i に出力する．

なお，本研究では，破綻した発話の検出は正確に行われた状況を仮定し，破綻検出は行わない．従って，破綻した発話の分類のみを行い，破綻ではない時刻 i の出力 $\mathbf{e_i}$ は 0 ベクトルとし，評価の対象としない．

3.3　各グループに対する手法

以下に，発話対象範囲軸を基にしたそれぞれのグループに対する提案手法について示す．

3.3.1　発話レベル

発話レベルでは，当該のシステム発話そのものに関するエラーが考慮される．

[発話-形式] グループは，日本語として妥当な文字列，または文構造を成していない発話をエラーとする．日本語として妥当ではない場合，対象となるシステム発話の単語間の繋がりが不自然になると考えられる．そこで，N-gram を用いた言語モデルを作成し，単語間の繋がりの不自然さを検出する．

言語モデルについては，Kneser-Ney による平滑化を利用したモデルを構築する．加えて，文法的な不自然さを適切に考慮する為，コーパス中の名詞を正規化し学習を行う．

[発話-内容] 軸は，単語の意味的組み合わせの誤用や，事実と異なる発話をエラーとする．意味的な組み合わせが不適切である場合，主語と述語，述語と目的語の組み合わせが相応しくないと考えられるため，格フレーム辞書，又は事前学習モデルのマスキングを用いて検出を行う．一方，事実と異なる発話の場合は，大規模な知識を必要とする為，ツイートの検索を利用した検出を行う．事実と大きく異なる内容はツイートされないという仮定と，事実と異なる発話は，固有名詞や希少な単語に関する発話である可能性が高いという仮定に基づき，検出を行う．

3.3.2　応答レベル

応答レベルでは，当該のシステム発話と，その直前のユーザ発話に関するエラーが考慮される．

[応答-形式] グループは，直前のユーザ発話が ”質問”，”依頼”，”提案”，”挨拶” を意図した前向き機能を持つ時に，その機能に対応した後ろ向き機能を有さない発話をエラーとする．これに対し，前向き機能を有する発話の検出器と，前向き機能に対応する後ろ向き機能を持つかを分類する分類器を構築し，それらを組み合わせることで検出を行う．

前向き機能については，文末が「〜ですか？」や「〜はいかがでしょう？」となる場合が多く，言語的特徴を有する．その為，n-gram 等の素性が検出に有効だと

考えられる．素性については，福岡ら [5] の対話行為推定の研究より，質問 (YesNo)，質問 (What)，要求の 3 つについて有効とされた素性を利用する．検出器の学習には，サポートベクタマシン (SVM) を利用する．一方，後ろ向き機能については,「はい」「いいですよ」等の短い応答については言語的特徴を有するが，質問に対して具体的な内容を返す応答の場合は，言語的特徴よりも発話内容の意味を考慮する必要があると考えられる．そこで，sentence BERT を用いて分類器を構築する．

[応答-内容] グループは，ユーザ発話に対する応答の形式は正しいが，期待された内容が含まれていないシステム発話をエラーとする．エラー例を見ると,「〜はありますか」の場合に,「はい」と答えた場合が期待を無視したとされているので，ルールベースと機械学習を併用した検出を検討している．

3.3.3 文脈レベル

文脈レベルでは，当該のシステム発話を含む，それまでの対話列全体に関するエラーが考慮される．

[文脈-形式] グループは，ユーザ発話の話題に対応していない発話や，システム発話が直前のユーザ発話の機能に対して不適切な機能を持つ発話などをエラーとする．話題遷移に関するエラーは，豊島ら [6] の発話ベクトルの差分を用いた手法を拡張し，検出を行う．一方，ユーザ発話の機能に対して不適切な機能を持つ発話の検出については，当該システム発話と，その発話から遡った 4 発話の 5 つの発話を入力とし，事前学習モデルを用いた検出を検討している．

[文脈-内容] グループは，過去の発話に対しての矛盾や，不要な繰り返しをする発話をエラーとする．繰り返しについては，過去の発話との N-gram の一致率，及びレーベンシュタイン距離の類似度，文章ベクトルの類似度に基づいた検出を行う．過去発話に対する矛盾の検出については，自然言語推論 (NLI) 技術を用いた検出を行う．NLI は「前提」と「仮説」からなる 2 つのテキストを用いて学習を行い，前提と仮説が推論可能か，矛盾しているか，中立かを予測するタスクである．前提を直前のユーザ発話までの対話，仮説を当該システム発話とすることで，矛盾の検出を試みる．

3.3.4 社会レベル

社会レベルでは，当該のシステム発話と一般的な社会性に関するエラーが考慮される．

[社会-形式] グループは，言葉遣いが礼節に反し，ユーザを不快にさせる，または差別的な発言を含むシステム発話をエラーとする．このグループについては，瀬川ら [7] の手法を利用する．具体的には，Twitter にクエリを掛けて攻撃的な内容を含むコンテンツを収集し，Sentence-BERT[8] を利用して検出することを検討している．

[社会-内容] グループは，社会通念と反するにも関わらず，根拠なく断定或いは肯定するシステム発話をエラーとする．対象とするデータセットでは,「熱中症は大丈夫ですね」「死者は良いですね」といった，明らかにネガティブな極性の単語に対し，ポジティブな極性の単語が形用されている場合が多い．堀井ら [3] の環境類型に対する手法を利用し，検出する．

3.4 マルチラベル分類の方法

本研究では，対話破綻とされたシステム発話を，8 グループへ分類する．マルチラベルに分類するにあたり，2 つの方法がある．

1. 各モジュールが独立に検出
2. 発話レベルモジュールの優先度を考慮

1. の場合，図 1 で示された 4 つのモジュールが出力した結果をそのまま統合し出力する．一方，2. の場合，もし発話レベルのエラー類型に分類された場合は，以降のモジュールの出力を 0 ベクトルとする．この優先度は，評価データセット中の規則に則ったものである．

本実験では，以上の 2 つの方法で結果を検討する．

4 進捗

現在，それぞれのグループに対して提案した手法を実装，評価中である．現時点での提案手法の結果を，表 4 に示す．

表 4 中で，空白はまだ提案手法が実装出来ていないことを表す．また，斜字の値については，実装途中の不完全な値を表す．

4.1 各グループに対する進捗

4.1.1 発話レベル

[発話-形式] グループについては，Keyser-Ney による言語モデルを実装した．言語モデルを学習するコーパスとして，学習データ中のユーザ発話と破綻ではないシステム発話を 2851 文，名大対話コーパス [10] から 2 者対話の発話文を抽出した 43385 文，sugiyama らの研究で構築された雑談対話コーパス [9] から 141777 文

表 2: 実験結果

	形式				意味			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
発話	0.915	0.033	0.571	0.064	*0.810*	*0.608*	*0.893*	*0.723*
応答	0.829	0.638	0.768	0.697				
文脈	*0.698*	*0.692*	*0.729*	*0.710*	*0.981*	*0667*	*0.536*	*0.594*
社会	*0.979*	*0.000*	*0.00*	*0.00*				

の計 188013 文を利用した．表 4 より， Recall が 0.6 に満たず，検出が出来ていない発話が多いことが得られた．

[発話-内容] グループについては，事実と異なる発話を検出した結果を示している．単語を形態素解析し，wikipedia から作成した idf 辞書を用いて，希少な単語を含むシステム発話を抽出した．抽出した発話を文節に区切り，文節を 2-gram で連結した各文字列をクエリとして，逐次 Twitter 検索をした．検索結果が表れなかったクエリを含むシステム発話を誤情報とし，このグループのエラー類型とした．表 4 より， Recall が 0.893 であり，事実と大きく異なる内容はツイートされないという仮定が概ね正しいことが得られた．一方，precision が 0.608 と高くないことから，日常的で一般的な内容であっても，全く同じ表現がツイートされるとは限らないことが得られた．

4.1.2 応答レベル

[応答-形式] グループについては，SVM によるユーザ発話の前向き機能検出器と，Sentence-BERT による分類器を組み合わせた結果を表している．表 4 より，Precision が 0.692 であり，精査するとシステム発話が誤情報の場合に誘発されていることが多いと判明した．今後は．システム発話側の抑制方法についても検討する．

[応答-内容] グループは，全く実装が進んでいない．また，学習用コーパス及び評価用コーパスにデータが存在しないため，精度による評価が出来ない．その為，人手で少数のエラーを作成し，別途評価する予定である．

4.1.3 文脈レベル

[文脈-内容] グループは，ペルソナチャットの対話を 1 発話ごと抜き出し，疑似的にエラーを作成した学習データによって sntence-BERT を学習した結果を示している．表 4 より，F1-score が 0.7 を超えているが，この疑似エラーの有効性を調査していない点があり，実装途中とする．また，このグループに属する話題遷移エラーについては，豊島らの手法を拡張した手法の適用を検討しているが，その未だ実装途中である．

[文脈-内容] グループは，不要な繰り返しを検出した結果を示している．現時点では n-gram の一致率を用いた結果である為，今後は文章ベクトル，レーベンシュタイン距離の実装を行い改善する．自然言語推論を学習する為のコーパスとして，機械翻訳を用いて用意された日本語 SNLI コーパス [11] を利用する．日本語 SNLI コーパスを sentence-BERT でファインチューニングし，矛盾した発話の検出を検討してているが，まだ実装途中である．

4.1.4 社会レベル

[社会-形式] グループについては， Twitter から，非常に強い暴言を 1138 件収集し，手動でラベリングを行った．ラベルは，「0:一般の発話，1:自己を卑下した発話 (攻撃的ではない), 2:第三者に対する攻撃, 3:対話相手 (ユーザ) に対する攻撃」と設定した．ラベリングの後に sentence-BERT によって学習を行ったが，表 4 より，評価データに対しては，現時点で有効な結果が得られなかった．これは，収集したコーパスの性質と，評価データ中にある [社会-形式] グループによるエラーの発話の傾向が異なっていた為だと考えられる為，改めてコーパスを収集することを検討している．

[社会-内容] グループについては，日本語評価極性辞書の用言編 [12], 名詞編 [13]，を用いた検出を検討しているが，まだ実装途中である．

5 今後の課題

- 全体として，比較する対象が必要のため，早急にベースラインを実装する
- 残りのグループについて実装する
- 全体を評価，考察する

参考文献

[1] 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. 第 75 回言語・音声理解と対話処理研究会 (第 6 回対話システムシンポジウム), 人工知能学会研究会資料 SIG-SLUD-75-B502, pp. 27-32(2015)

[2] Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara and Masahiro Mizukami：Integrated taxonomy of errors in chat-oriented dialogue systems, SIGDIAL 2021.

[3] 堀井朋, 森秀晃, 林卓也, 荒木雅弘: 破綻類型情報に基づく雑談対話破綻検出, 言語・音声理解と対話処理研究会, vol.78, pp.75-80, 2016

[4] 稲葉 通将, 高橋 健一: 対話破綻検出による対話システムの応答性能の向上, 人工知能学会研究会資料, SIG-SLUD-B508-29, pp.110-115

[5] 福岡 知隆, 白井 清昭: 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, 言語処理学会 第 22 回年次大会, 2016, pp.1121-1124

[6] 豊嶋 章宏, 吉野 幸一郎, 須藤 克仁, 中村 哲 : 発話ベクトルの差分特徴量を用いた雑談対話システムにおける破綻した話題遷移の検出, 言語処理学会 第 24 回年次大会, p873-p873

[7] 瀬川 友香, 浅谷 公威, 坂田 一郎: ユーザーに着目した SNS 上の攻撃とそのメカニズムに関する分析, 2021 年度人工知能学会全国大会（第 35 回） pp.1-4

[8] Reimers, N. and Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019)

[9] Hiroaki Sugiyama and Masahiro Mizukami and Tsunehiro Arimoto and Hiromi Narimatsu and Yuya Chiba and Hideharu Nakajima and Toyomi Meguro: Empirical Analysis of Training Strategies of Transformer-based Japanese Chitchat Systems, arXiv, 2109.05217, 2017

[10] 藤村逸子・大曽美恵子・大島ディヴィッド義和、2011「会話コーパスの構築によるコミュニケーション研究」藤村逸子、滝沢直宏編『言語研究の技法：データの収集と分析』p. 43-72、ひつじ書房

[11] 吉越 卓見, 河原 大輔, 黒橋 禎夫: 機械翻訳を用いた自然言語推論データセットの多言語化, 第 244 回自然言語処理研究会, (2020.7.3).

[12] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. 意見抽出のための評価表現の収集. 自然言語処理, Vol.12, No.3, pp.203-222, 2005.

[13] 東山昌彦, 乾健太郎, 松本裕治, 述語の選択選好性に着目した名詞評価極性の獲得, 言語処理学会第 14 回年次大会論文集, pp.584-587, 2008.