# 対話破綻検出を用いた雑談対話の精度向上

Improvement of Chat Dialogue Using Dialogue Breakdown Detection

高橋 龍平 [1)]
指導教員　岩下 志乃 [1)]

1）東京工科大学大学院 バイオ・情報メディア研究科 コンピュータサイエンス専攻 岩下研究室

キーワード：チャットボット，対話破綻，対話システム，雑談対話

## 1．はじめに

近年では,端末技術の進歩や AI の研究の発展により，AI を搭載したロボットやスマートスピーカーが増加している．非タスク指向型対話システムでは雑談対話が可能で，Romi や Charlie などは人とコミュニケーションを図れる．しかし，現在の雑談対話システムでは対話が破綻してしまう可能性がある．星野ら[1]によると,人は雑談を通して相手の印象を形成したり，人間関係を調整したりしている．そのため，対話システムの対話破綻を減らすことが出来れば人が親しみやすくなり，より人間らしい雑談対話システムの実現に繋がると考えられる．

対話が破綻しないために，様々な研究が行われている.対話破綻検出チャレンジ(DBDC)[2]の結果から,対話破綻を事前に検出することが可能であることが分かっている．また，稲葉ら[3]の研究では，対話システムの応答に対し，対話破綻検出手法を適用することで，対話の自然さが向上していた．

本研究では，非タスク指向型対話システムに対話破綻検出器を用いて対話精度の向上と対話システムの印象向上を図る．また，対話破綻検出器に有用と考えられる破綻の特徴の抽出を行う．

## 2．方法

### 2.1．システムの概要

今回作成するシステムの処理の流れを図 1 に示す．雑談対話システムで応答候補を生成し，その応答候補に対して対話破綻検出器で応答が破綻しているかを確認する．破綻していると判断された場合は，破綻の特徴に合わせた新たな応答を生成する．

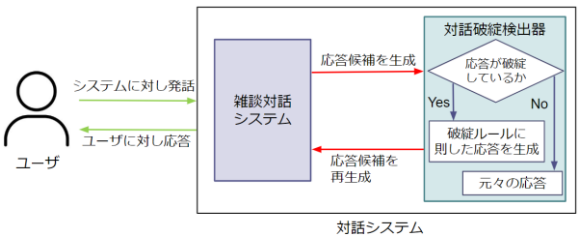


図 1　提案手法の処理の流れ

### 2.2．対話システムの作成

雑談対話システムを作成するにあたって，学習に用いるデータは，Twitter から取得したツイートとそのツイートに対するリプライを 1 つの対話データとする．取得した「ツイート＋リプライ」のペアのデータから，個人名と外国語，メンション，絵文字，顔文字等の学習に影響を及ぼすと考えられるものを削除した．学習には，OpenNMT というオープンソースのニューラル機械翻訳フレームワークを使用する．

### 2.3．検出する破綻の特徴

対話破綻検出器で検出する破綻の特徴を以下に示す．

- 話題への固執，不必要な繰り返し：システムが特定の話題に固執し，ほぼ同じ内容の発話を繰り返す
- 長すぎる発話：システムの発話が長いほど無関係な部分が含まれる可能性が高くなる
- 経過ターン：対話ターンが経過するごとに不適切な発話の割合が増えていく
- 発言として唐突：挨拶に関係ない発言を唐突に行う
- 疑問文に対する応答：Twitter のデータを使用しているため，システムの自己開示に破綻が出る可能性がある

杉山の手法[4]で提案されている誤りパターンと，作成したシステムの破綻になりそうな箇所をピックアップして破綻の特徴を決めた．破綻の特徴はルールベースで検出を行う．

## 3．評価実験と考察

### 3.1．実験内容

実験協力者 13 名に対話破綻検出器を実装していない対話システム(System1)と,対話破綻検出器を実装した対話システム(System2)の両方と対話を行っ

てもらう．実験協力者にはどちらのシステムが対話破綻検出器を実装したシステムであるかは伝えていない．実験後のアンケートは，対話システムの「自然さ，楽しさ，総合(優れている)」の 3 つについて，どちらのシステムが良かったかを 5 段階の評価で選んでもらった．

また，実験協力者と対話システムとの対話ログを用いて，アノテーターに破綻ラベル付けを行ってもらった．破綻ラベルは以下の 3 種類である．

○：破綻ではない

△：破綻とは言い切れないが，違和感のある発話

×：明らかにおかしいと思う発話，破綻

### 3．2．実験後アンケートの結果と考察

実験後のアンケートの結果を図 2 に示す．図 2 の結果から，提案手法である対話破綻検出器ありの方がユーザに与える印象が良くなっていることが分かる．

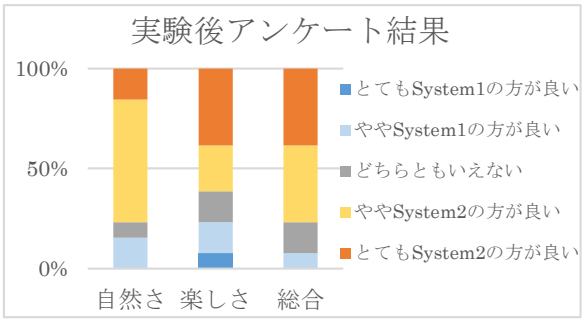


図 2　実験後アンケートの結果

### 3．3．アノテーターの評価結果と考察

破綻ラベルを付けるアノテーターは 6 名であり，System1 と System2 に対して各アノテーターが「○」と評価した割合の比較結果を表 1 に示す．破綻の割合には，System1 と System2 に大きな差は見られなかった．提案システムの方がアンケートの結果が高かった理由は破綻が減ったからとは言えない．

表 1　各システムに対して各アノテーターが「○」と評価した割合の比較

| アノテーター | System1 | System2 |
|---|---|---|
| Annotator_1 | 0.4864 | 0.5814 |
| Annotator_2 | 0.5429 | 0.5406 |
| Annotator_3 | 0.5225 | 0.5484 |
| Annotator_4 | 0.5918 | 0.6243 |
| Annotator_5 | 0.6220 | 0.6499 |
| Annotator_6 | 0.4557 | 0.5004 |
| 平均 | 0.5369 | 0.5742 |

### 3．4．対話破綻検出器の検出結果と考察

表 2 に各破綻特徴の検出回数を示す．「不必要な繰り返し」と「発言として唐突」は，検出回数は少ないが「○」の評価が 100%であったため破綻の特徴として有用であったといえる．「疑問文に対する応答」では，対話破綻検出器が精度向上に有用に働いた場合と，反対に精度を下げてしまうという場合の両方が見られたが，検出回数の多さから有用であると考えられる．しかし，「長すぎる発話」に関しては検出回数が 0 回であるため，今回のシステムでは有用に働かなかった．

表 2　各破綻特徴の検出回数

| 破綻の特徴 | 検出回数 | ○ | △ | × |
|---|---|---|---|---|
| (破綻ではない) | 109 | 64 | 20 | 25 |
| 不必要な繰り返し | 3 | 3 | 0 | 0 |
| 長すぎる発話 | 0 | 0 | 0 | 0 |
| 経過ターン | 8 | 1 | 7 | 0 |
| 発言として唐突 | 11 | 11 | 0 | 0 |
| 疑問文に対する応答 | 64 | 13 | 3 | 48 |

### 4．おわりに

本研究では，対話破綻検出器により検出した破綻を修正することで，対話精度の向上と対話システムの印象向上を図り，対話破綻検出器に有用と考えられる破綻の特徴の抽出を行った．実験後のアンケートでは，対話破綻検出器ありの対話システムの方が評価は高かったが，アノテーターの評価結果では，破綻の数に大きな差はなかった．

今後の課題は，検出する破綻の見直しである．特に疑問文への対応では，類似文を検索する等の新たな検索方法を検討する必要がある．

### 参考文献

[1] 星野春香，松本知香，“心理臨床面接における雑談の可能性についての一考察”， 京都大学大学院教育学研究科附属臨床教育実践研究センター紀要，Vol.24，pp.110-117，2021.

[2] 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”， 人工知能学会資料，SIG-SLUD-B502-07，pp.27-32，2015.

[3] 稲葉通将，高橋健一，“対話破綻検出の対話システムへの適用”， 人工知能学会論文誌，Vol.34，No.3，pp.1-8，2019.

[4] 杉山弘晃，“発話生成における誤りパターンの分析に基づく対話破綻検出”， 人工知能学会研究会資料，SIG-SLUD-B505-23，pp.81-84，2016.