E-008

駄洒落を用いた雑談対話システムにおける対話破綻回避の有効性

Effectiveness of Japanese Pun to Avoid Dialogue Breakdown on Chat Dialogue System

徐　云帆† 荒木　健治†

Yunfan Xu Kenji Araki

1. はじめに

雑談対話中では多様な話題を扱う必要があるため，雑談対話システムは適切な応答を返すことに様々な困難を抱える．それゆえ，雑談対話システムは未だに多くの対話破綻が存在する．このような対話破綻を回避することにより雑談対話を高精度化することが可能になる．対話破綻が起こる原因はシステム発話とユーザの期待の乖離である．また，ユーモアの研究において，ユーモアの生起には「不適合理論」という理論がある．大部分のユーモア研究者が，ユーモアの生起に不適合の認知が不可欠であるという点で合意を達成しつつあると言われる[1]．研究者は期待と実際のズレからユーモアが生まれると考えている．ユーモアの生起と対話破綻が発生する要因は共通点があるため，ユーモアを用いることにより，対話破綻を回避できる可能性がある．

本研究では対話破綻をユーモアにより回避するという立場から破綻がある応答文を駄洒落ユーモアで置き換えるという破綻を回避する手法を提案し，その有効性を考察する．

対話破綻について，破綻検出チャレンジ[2]が開催されており，最近では機械学習手法の改善やディープラーニングを用いることにより，検出精度が向上している．しかし，対話破綻を検出した後の処理として破綻回避に関する研究は未だに少ない．駄洒落について，谷津ら[3]は駄洒落ユーモアの生成及び認識の双方を適切な頻度において行う統合型対話システムを開発しているが，対話破綻の問題点については扱っていない．

本稿では，評価実験により駄洒落を用いて回避可能な対話破綻を考察し，それに関するアノテーションを行い，コーパスを構築した結果について述べる．

2. 評価実験

対話破綻回避に駄洒落を利用することの有効性と回避可能性に影響する要因を考察するために，評価実験を行った．

2.1 実験概要

破綻検出システムを用いて破綻検出を行い，破綻がある箇所のユーザ発話を駄洒落生成システムに入力する．生成された結果と元の破綻文を比較し，人手による評価を行った．

2.2 実験条件

実験に使用したデータは文献[2]の破綻検出チャレンジの評価データで，そのデータは NTT ドコモが公開している雑談対話 API を用いて収集された対話データである．破綻検出は久保ら[4]の Neural Conversational Model を用いる深層学習に基づく手法を基に開発した．この手法は対話破綻チャレンジ 2 において比較的 F 値の数値が高い手法である．駄洒落生成には図 1 で示したように，(1)入力文の改変（漫才のボケ形式）(2)なぞなぞ(3)駄洒落データベースのいずれかのタイプの駄洒落を生成するシステム[5]を用いる．

破綻検出モジュールにより検出された 376 文の破綻文からランダムに 50 文を抽出し，それに対応するユーザ発話を駄洒落生成システムに入力する．生成した駄洒落文と元の破綻文を 1 点(悪い)から 5 点(良い)の五段階で人手により評価した．評価者は 20 代理系男性大学院生 6 名である．

(1)入力文の改変
入力：現代語訳されていれば簡単ですよ，内容がわかりにくいことはありますが.
出力：「最近，現代語訳されていれば簡単ですよ，内容がわかりにくいことはありますがっていう感じでしたよ」「俺の場合，現代語訳されていれば天丼ですよ，ナチョスがわかりにくいことはありますがって感じだけど」「寒い！」
(2)なぞなぞ
入力：海外の女性はキャリアウーマンってイメージあるね.
出力：「海外は海外だけど，食べれる海外はな〜んだ？」「何？」「海貝」
(3)駄洒落データベース
入力：何事は行っていませんが，明日の予定を考えています.
出力：予定と言えば，今日の予定は，飲み会のみかいって言いたくなるよね.

図 1　駄洒落生成の例

2.3 実験結果

表 1　評価実験の結果

	平均値	平均二乗誤差
破綻文	2.023	0.528
駄洒落文	**2.040**	**0.208**

評価実験の結果を表 1 に示す．表 1 で駄洒落文は平均値で 0.017 ポイント破綻文を上回った．破綻文の平均二乗誤差が高い原因は破綻検出モジュールの検出精度が 40%でまだ不完全なため，正しい文が破綻として検出された場合，評価に影響を及ぼしたためである．

表 2　駄洒落のタイプごとの実験結果

	文数	平均値	平均二乗誤差
(1)入力文改変	2	1.833	0.0
(2)なぞなぞ	21	2.024	0.212
(3)データベース	27	**2.068**	0.217

† 北海道大学大学院　情報科学研究科, Graduate School of Information Science and Technology, Hokkaido University

表 2 に示すように，駄洒落データベースの駄洒落が最も良い結果となった．(1)入力文改変と(2)なぞなぞの評価が低い原因は駄洒落自体が会話文のため，一つの発話として違和感があったためである．

2.4 考察

実験結果から見ると，評価の良い例と評価の悪い例それぞれ特徴がある．評価の良い例では発話が平叙文の場合が多く，駄洒落生成システムがヒットする単語の多数は普通名詞である．一方，疑問文や具体的な要求のある文に対し，駄洒落の回避が不自然と評価されることが多い．

駄洒落を用いた対話破綻回避が有効になる対象文は限定的であり，影響する要因は多様である．対話破綻回避の精度を上げるため，回避可能な破綻を検出する必要がある．したがって，今後機械学習により駄洒落を用いた回避可能な破綻を検出し，特定の対話破綻を駄洒落生成モジュールによりリカバリする手法を提案する予定である．そのために学習に必要となるコーパスを作成した結果について述べる．

3. 回避可能破綻コーパス

破綻ラベル付きの雑談対話コーパス[6]を基づき，対話破綻の各箇所の駄洒落使用の妥当性を評価しアノテーションを行い，コーパスを作成した．

3.1 アノテーションの方法

破綻ラベルが付いた破綻文直前のユーザ発話に対し駄洒落生成を行い，生成した駄洒落を参考として，文脈から駄洒落の使用が自然か否かを判断する．

アノテーションに使用した駄洒落生成システムは，新しい駄洒落データベースにより作成した駄洒落生成システム[7]である．前章の実験結果を踏まえた上で，ヒットする単語の優先度を設定した．発話に対し MeCab を用いて形態素解析を行い，「一般名詞・固有名詞」が含まれる場合は，ランダムに一つを選び，駄洒落生成システムの入力とする．含まれない場合は，「サ変名詞・動詞・形容動詞・形容詞」の中から単語を一つ抽出し駄洒落生成を行う．両方とも含まれない場合は一つ前の発話文を対象に変換する．

3.2 予備的アノテーション

3.1 で説明したアノテーション方法では，アノテータ間の個人差や生成した駄洒落の質がアノテーションに影響すると考えられる．これらの要素がもたらす影響を考察するため，予備的アノテーションを行った．

3.2.1 実験条件

対話破綻コーパス中からランダムに 20 対話を抽出し，**3.1** で説明した方法により駄洒落を生成し，応答文に付与する．前半の 10 対話は個人差の評価に使用するため同じ駄洒落を付与する．後半の 10 対話は駄洒落の質による影響を考察するため，付与する駄洒落を 3 種類の異なるものとした． 3種類のアンケートを用いて7人のアノテータ(理系大学院生男性 6名女性 1名)によりアノテーションを行った．

3.2.2 実験結果及び考察

アノテータの一致度を考察するためにフライスの κ 値を算出した．前半の 10 対話の結果から算出したフライスの κ 値は 0.17 で，後半のフライスの κ 値は 0.24 である．

後半は前半より一致度が高いため，駄洒落の質により評価への影響は少ないと考えられる．全体のフライスの κ 値は 0.2 程度なので，「ランダムではないが，おおむね一致」ということである[8]．

予備的アノテーション時のアノテータのコメントにより「話題遷移のところの駄洒落の効果が高い」「会話の冒頭に駄洒落を使うのは不自然」「疑問，要求の後に駄洒落を使うのは不自然」などの意見があった．

3.3 コーパス作成

今回収集したコーパスは合計 350 対話である．その中で破綻文数は 799 文で，アノテーションによる回避可能な破綻文数は 259 文である．

4. まとめと今後の課題

本稿では，駄洒落ユーモアにより対話破綻を回避する手法の有効性を考察し，それに関するコーパス構築を行った．評価実験の結果より，提案手法がある程度有効であることを確認されたが，さらに精度を上げるには回避条件を絞る必要がある．駄洒落使用の主観性が高いため，アノテーション方法でヒットする単語を調整することや一致度を計算することにより，より客観的なデータが得られた．また，今回収集したコーパスのデータ量は比較的少ないため，学習する際に使用する手法を検討する必要がある．

今後の課題としては，回避可能破綻コーパスを学習することで回避可能な破綻を検出することやシステム発話をより自然にするため破綻回避に必要な駄洒落の表現手法などが挙げられる．

謝辞

本研究は科研費(基盤研究 (C)17K00294)の助成を受けたものである．

参考文献

[1] 伊藤 大幸, “ユーモアの生起過程における論理的不適合および構造的不適合の役割”, 認知科学, Vol.17(2) No. 297-312 (2010).
[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”, SIG-SLUD, Vol.5(02),No. 27 - 32 (2015).
[3] 谷津元樹, 荒木健治, “話題遷移に適応した駄洒落ユーモア統合型対話システムの性能評価”,ことば工学研究会: 人工知能学会第 2 種研究会ことば工学研究会資料,Vol.52,No. 23-27. (2016)
[4] 久保隆宏, 中山光樹“Neural Conversational Model を用いた対話と破綻の同時学習”. SIG-SLUD, Vol.5(02), No.94-97(2016).
[5] Jonas Sjobergh, Kenji Araki, “Evaluation of a Humor GenerationSystem by Real World Application with ￥500,000 to Win. In Proceedings of the Linguistic And Cognitive Approaches To Dialog AgentsSymposium”, (2010).
[6] 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(1), No.59-86. (2016).
[7] 荒木健治, “駄洒落データベースを用いた駄洒落生成システムの性能評価”, 人工知能学会第 2 種研究会:ことば工学研究会資料, SIG-LSE-B Vol.703-8, No.39-48, (2018).
[8] Landis, J. R., Koch, G. G, “The measurement of observer agreement for categorical data. Biometrics”, 159-174(1977).