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

Chat-oriented Dialogue Breakdown Detection based on the Analysis of Error Patterns in Utterance Generation

杉山 弘晃 [1]
* Hiroaki Sugiyama [1]

[1] NTT コミュニケーション科学基礎研究所
[1] NTT Communication Science Laboratories

Abstract:
Chat-oriented dialogue systems sometimes generate utterances that are inappropriate as the responses for user utterances and cause dialogue breakdown. Some utterance generation errors have typical patterns such as inconsistent dialogue acts or dialogue topics. If a system detects these error patterns, it helps to continue dialogue with suppressing such inappropriate system utterances. In this paper, I develop a dialogue breakingdown detector and analyze the effects of training features, data and algorithms for dialogue breakdown detection performance.

1 序論

近年，従来のタスク指向の対話システムとは異なる，雑談を行う対話システムに注目が集まっている [大西 14, Ritter 11, Wong 12, 東中 14]．雑談対話は，エンタテインメントやカウンセリング目的のみならず，ユーザの潜在的な要求を引き出したり，ユーザと良好な関係を構築する上で重要である．

雑談対話システムは，ユーザ発話に含まれる非常に幅広い話題に応答する必要がある．そのため，適切な応答を出力し続けることは難しく，現在の対話システムでは，対話を破綻させるような発話がしばしば生成される．こうした破綻する可能性のある発話を予め検出し，出力を抑制することができれば，対話の継続が容易になると考えられる．

一方，現在の対話システムでは，システム発話を出力する際，何らかの構造に当てはめる生成ベースのアプローチや，あらかじめ記述された文を検索する検索ベースのアプローチが用いられることが多い．また，用いたアプローチに応じて，しばしば特定の誤りのパターンが観察される．例えば，生成ベースのアプローチでは，ユーザ発話に関連する話題を含むシステム発話を生成しやすい一方，テンプレートに想定外の単語が当てはめられてしまい，文として破綻しているパターンがある．また，検索ベースのアプローチでは，非文は生成されにくいものの，ユーザ発話に適切に合致する話題のシステム発話を検索できず，話題が合致しないシステム発話が生成されやすい特徴がある．こうしたアプローチごとの誤りのパターンをうまく検出することができれば，着実に発話の品質改善につながると期待できる．本研究では，対話破綻検出チャレンジ 2 で東中らが提供しているデータセット [東中 16] を用いて，上記誤りパターンの分析と，破綻検出に有用な特徴量の検証を行う．

*連絡先：NTT コミュニケーション科学基礎研究所
〒 619-0237 京都府相楽郡精華町光台 2-4
E-mail: sugiyama.hiroaki@lab.ntt.co.jp

2 比較要素

本節では，比較に用いた特徴量，訓練データ，およびアルゴリズムの詳細を説明する．本研究で分析の対象とする，対話破綻検出チャレンジ 2 で配布されているデータには，以下の 3 つのシステムのいずれかとユーザとの 2 者間テキストチャット，および各システム発話に 30 名が付与した評価アノテーション（○，△，×＝O, T, X）が収録されている．

DCM NTT ドコモによって提供されている，雑談対話 API.

DIT デンソー IT ラボラトリによって提供された．単語と対話パターンの共起グラフ上のラベル伝播による応答文生成 [Tsukahara 15] と，対話行為・ト

ピックによる応答選択 [塚原 16] を行った対話システム．

IRS オーガナイザが準備した，Ritter らによる IR-status と同等の用例ベース対話システム [Ritter 11]．

実際の出力例を観察したところ，このうち DCM は，汎用的なテンプレートに，ユーザ発話に合致する話題をはめ込んで発話を生成する生成ベースアプローチ，DIT と IRS は人が作成した用例をそのまま出力する検索ベースのアプローチが用いられていた．

2.1 誤りパターンと対応する特徴量

本研究ではまず，各システムに特有の誤りパターンを分析し，個々に有用と考えられる特徴量を提案する．

急激な話題転換：単語一致率 特に DIT と IRS において，ユーザ発話と全く異なる話題の発話が生成されるケースが散見された．これらは用例ベースのシステムであるため，適切にユーザ発話の内容に合致する用例が見つからない，もしくは存在しない場合でも，与えられた尺度の上で最も類似した用例が選択されてしまう問題がある．本研究では，破綻検出の対象発話と，その直前のユーザ発話との内容語の単語一致率を，急激な話題転換を検出する特徴量として用いる．

話題への固執・不要な繰り返し：単語一致率 特に DCM において，システムが特定の話題に固執し，結果としてほぼ同じ内容の発話を繰り返す場合が多く見られた．DCM においても，まったく同じ文字列の発話は抑制されていると思われるが，表記のゆれや語尾の変化などにより，フィルタリングが不十分なものと考えられる．本研究では，破綻検出の対象発話と，その 2 発話前の発話（直前のシステム発話）との内容語の単語一致率を，話題への固執・不要な繰り返しを検出する特徴量として用いる．

対話行為の不自然なつながり：対話行為，文字列共起，述語 各システムを通して，ユーザ発話が質問の場合でも，システムが質問で応答するケースが多く見られた．またそもそも，今回のシステムでは，質問に答える機能が不足しているように感じられた．これらを検出するため，本研究では，破綻検出の対象発話と，その直前のユーザ発話について，その発話が表す対話行為，およびその次発話としてふさわしいと予想される対話行為の 2 種類を推定し，特徴量として加える．

上記推定器は，別途 NTT の雑談対話コーパス（3680 対話）[Higashinaka 14] から，線形 SVM・単語 1,2gram 特徴を用いて学習した [Sugiyama 13]．また合わせて，対象発話・直前のユーザ発話間での文字 4gram の共起（頻度 4 以上）と，対象発話・直前のユーザ発話に含まれる述語も特徴量に加える．

長すぎる発話：文長 現在の対話システムの技術では，ユーザ発話の内容とシステム発話の内容との一貫性を誤りなく推定することは困難である．そのため，システム発話が長ければ長いほど，無関係な部分が含まれる可能性が多くなってしまうという問題がある．今回の 3 システムの中では，DIT のシステム発話が全体的に非常に長く，それゆえユーザ発話の内容に合致しない発話となっているケースが多く見られた．本研究では，破綻検出の対象発話の単語長および文字長を特徴量に加える．

シナリオ内：頻出単語列 DIT において，A と B のどちらの話をするか，のような，あらかじめ想定されたシナリオに誘導するシステム発話が観察された．この直後では，比較的それまでの文脈から切り離されてシステムが応答できるため，用例ベースにとって有利な条件であると考えられる．しかしながら，実際の発話例を見てみると，必ずしも自然な流れではない場合もあった．また，実際に対話破綻を推定することを考えると，シナリオの直後では，他の部分とは評価傾向が異なると想定される．そのため，シナリオ直後か否かを推定するための特徴として，頻出する単語 6gram の文字列（頻度 10 以上）を特徴量に加える．

経過ターン：ターン数 いずれの対話システムにおいても，対話の冒頭部分では比較的適切な発話を生成しているものの，対話が経過するごとに不適切な発話の割合が増えていく傾向がみられた．そのため，対話開始からの経過ターン数を特徴量に加える．

2.2 利用データ

今回のタスクでは，評価分布の一致に重きを置かれている．分布の一致は単純な F 値評価よりデータの性質に影響を受けやすいと予想される．実際に予備実験で前回の対話破綻検出チャレンジの評価データ（eval）を対象に推定した場合，最も類似する dev のみで学習した場合が最も性能が高く，類似のデータ（前回の対話破綻検出チャレンジでの rest1046，init100）を追加し

表 1: DCM，DIT，IRS に対する推定性能

抜いた特徴量	Accuracy	Mean squared error	JS divergence
マジョリティベースライン	0.447	0.0735	0.142
全特徴量	0.580	0.0443	0.0838
-ターン数	0.585	0.0455	0.0860
-対話行為	0.554	0.0537	0.0960
-単語一致率	0.571	0.0456	0.0859
-文長	0.582	0.0446	0.0843
-文字列共起	0.580	0.0443	0..0838
-頻出単語列	0.603	0.0405	0.0769
-述語	0.575	0.0449	0.0851
-(頻出単語列，述語)	**0.613**	**0.0402**	**0.0768**
-(頻出単語列，文字列共起)	0.599	0.0405	0.0770
-(頻出単語列，述語，文字列共起)	0.610	0.0405	0.0774

たところ，大きく性能低下していた．そのため本研究では，追加の学習データは用いず，配布された DCM, IRS, DIT のみを学習に用いる．

2.3 アルゴリズム

前節で説明した各特徴量は，個々に独立に利用するよりも，組み合わせを考慮できるほうが望ましい．一方，本研究では配布データのみを学習に用いるため，データ量には限りがある．また，本タスクでは，正解との○・△・×（= O, T, X）の分布間距離（Mean squared error，JS divergence）に重きが置かれているため，分布間距離を最小化するよう，回帰モデルを用いることが望ましい．そのため本研究では，RandomForest の派生形である，ExtraTreesRegressor を推定アルゴリズムに用いる．決定木同様，複数の特徴量の組み合わせを考慮した推定が可能である．なお，ExtraTreesRegressor の実装は scikit-learn[1] を利用する．

3 実験

3.1 実験設定

前章で説明した特徴量を組み合わせ，破綻検出器を構成し，破綻検出性能を比較する．本タスクでは，正解との○・△・×（= O, T, X）の分布間距離（Mean squared error，JS divergence）に重きが置かれているため，分布間距離を最小化する特徴量を検証する．なお，本研究では，全特徴量を用いた場合からいくつか特徴量を引くことで，その特徴量の重みを推定することとする．また，全ての学習データの OTX の分布を平均したものを，マジョリティベースラインとして用いる．実験では，DCM, DIT, IRS のデータを 3:1 にランダムに分割し，3/4 を学習データ，1/4 を評価データとして用いた．

3.2 結果

結果を表 1 に示す．全特徴量を用いたときの値を基準に，それよりも良い値（Accuracy では高い，分布間距離では低い）であれば，そのとき抜いた特徴量は性能を悪化させる特徴であり，逆に全特徴量よりも悪い（Accuracy では低い，分布間距離では高い）値であるほど，そのとき抜いた特徴量が推定に重要な役割を果たしているといえる．

表から，シナリオ内かどうかを判別するための頻出単語列特徴が，大きく性能を悪化させていることが分かる．学習データと評価データで，シナリオ後の対話破綻の度合いが大きく異なっていたものと考えられる．

一方，対話行為，単語一致率は性能向上に大きく寄与しており，発話の一貫性を表現するうえで重要な特徴であることがわかる．他の特徴は，おおむね Accuracy にはあまり影響がなく，分布間距離を多少改善する程度に留まっている．

上記をふまえ，頻出単語列を抜き，その上で分布間距離への影響が小さい述語，文字列共起，その両者を抜いて再度検証したところ，頻出単語列と述語を抜いた場合が，最も良い性能を示していた．今回のデータでは，頻度で足切りをしたとしても，述語のように次元が大きい特徴量は不向きであるといえる．

[1] http://scikit-learn.org/

表 2: 評価データに対する推定性能

提出 Run	Acc	MSE	JSD	f(X)	f(X+T)
提出 1	0.581	0.0411	0.0775	0.662	0.828
提出 2	**0.601**	**0.0410**	0.0775	**0.672**	0.809
提出 3	0.582	0.0411	**0.0773**	0.665	**0.829**

3.3 チャレンジタスクへの提出データ

チャレンジタスクへ提出したデータは，Accuracy，分布間距離の性能のよいものを中心に，試験的に学習データを追加したものを含めて，以下の 3 つを提出した．

提出 1 シナリオ内, 述語, 文字共起を除いて構築したモデル．開発データ上の Acc=0.613, MSE=0.0402, JSD=0.0768

提出 2 シナリオ内, 述語, 文字共起を除いて構築したモデル．学習データに，前回の破綻検出チャレンジの dev と eval を追加．開発データ上の Acc=0.610, MSE=0.0417, JSD=0.0794

提出 3 シナリオ内, 述語を除いて構築したモデル．開発データ上の Acc=0.610, MSE=0.0405, JSD=0.0770

評価データに対する結果を表 2 に示す．どれもほぼ同等の値ではあるものの，Accuracy，MSE，f (X) では提出 2 が，JSD と f(X+T) では提出 3 が最も高かった．ほぼ同じ傾向のはずの開発セットと異なる傾向を示していたことは興味深い．

4 結論

本稿では，雑談対話の破綻検出に有用な特徴量について分析した．分析の結果，学習対象のデータが少ない場合，単語のように次元が大きくなりやすい特徴量はやや不利であり，むしろ単語の一致率や対話行為など，従来より用いられている特徴量でも十分効果を発揮することが示された．一方，データ量を増やした場合には，こうした特徴量や，DNN のような，より抽象度の高い組み合わせ特徴を扱えるアルゴリズムを利用できる可能性がある．

課題として，今回導入した特徴量では，依然どのような話題展開が自然かについて，十分には扱えていない問題がある．ユーザ発話とシステム発話で単語が異なるとまとめて破綻とみなす傾向にあるため，話題が適切に展開していても，それを非破綻と評価できていない．これについては，大量のデータを援用し，話題の展開パターンを別途計算することで，ある程度補えると考えられる．大規模データを単純に学習データに追加しても性能はなかなか向上しないが，こうした例のように適切に情報を抽出して利用することで，破綻検出性能を向上させていきたい．

参考文献

[Higashinaka 14] Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., and Matsuo, Y.: Towards an open-domain conversational system fully based on natural language processing, in *Proceedings of the 25th International Conference on Computational Linguistics*, pp. 928–939 (2014)

[Ritter 11] Ritter, A., Cherry, C., and Dolan, W. B.: Data-Driven Response Generation in Social Media, in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 583–593 (2011)

[Sugiyama 13] Sugiyama, H., Meguro, T., Higashinaka, R., and Minami, Y.: Open-domain Utterance Generation for Conversational Dialogue Systems using Web-scale Dependency Structures, in *Proceedings of the 14th annual SIGdial Meeting on Discourse and Dialogue*, pp. 334–338 (2013)

[Tsukahara 15] Tsukahara, H. and Uchiumi, K.: System utterance generation by label propagation over association graph of words and utterance patterns for open-domain dialogue systems, in *Proceedings of Pacific Asia Conference on Language, Information and Computation* (2015)

[Wong 12] Wong, W., Cavedon, L., Thangarajah, J., and Padgham, L.: Strategies for Mixed-Initiative Conversation Management using Question-Answer Pairs, in *Proceedings of the 24th International Conference on Computational Linguistics*, pp. 2821–2834 (2012)

[大西 14] 大西可奈子, 吉村健：コンピュータとの自然な会話を実現する雑談対話技術, NTT DoCoMo テクニカル・ジャーナル, Vol. 21, No. 4, pp. 17–21 (2014)

[塚原 16] 塚原裕史, 内海慶：対話行為と話題推定によるラベル伝搬を利用した雑談生成方法の改良, 人工知能学会全国大会 (2016)

[東中 14] 東中竜一郎：雑談対話システムに向けた取り組み, 第 70 回言語・音声理解と対話処理研究会（SIG-SLUD）(2014)

[東中 16] 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, 第 78 回言語・音声理解と対話処理研究会（第 7 回対話システムシンポジウム）(2016)