The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

なりきりQAデータを用いた用例の拡張

Example Extension via Role-playing QA Data

水上 雅博 *1 Masahiro Mizukami
東中 竜一郎 *1*2 Ryuichiro Higashinaka
川端 秀寿 *3 Hidetoshi Kawabata
山口 絵美 *3 Emi Yamaguchi
安達 敬武 *3 Noritake Adachi
杉山 弘晃 *1 Hiroaki Sugiyama

*1NTT コミュニケーション科学基礎研究所 NTT Communication Science Laboratories
*2NTT メディアインテリジェンス研究所 NTT Media Intelligence Laboratories
*3株式会社ドワンゴ DWANGO Co., Ltd.

Data collection is one of the essential tasks of non-task-oriented dialogue systems. Role-playing question answering is proposed as an effective data collection approach to collect a large amount of consistent QA data at low cost. This QA data can use for example-based non-task-oriented dialogue system practically. However, collecting not only role-play QA data but also more diverse examples is important to answer various users ' questions. We propose an extension technique that increases the number of examples for non-task-oriented dialogue systems. This technique helps to erect non-task-oriented dialogue system, which can reply to question that does not include in role-playing QA data. In this research, we make a large number of consistent examples using a small amount of role-playing QA data and a large amount of Twitter data.

1. はじめに

従来のタスク指向の対話とは異なり，対話そのものを目的とする雑談対話システムの需要が高まっている [大西 14, Wong 12, Ritter 11]．雑談対話システムは，単に人間の対話相手になるのみでなく，エンタテインメントやカウンセリングへの応用や，対話を通したユーザ・システム間の信頼関係の構築への貢献も期待されている．

雑談対話システムを構築する手法の一つに用例ベース対話システム [Murao 03] の枠組みがある．この手法は，発話文や質問文と応答文を対にした用例と呼ばれるデータを利用し，ユーザの入力に対して最も近い発話文を持つ用例の応答文をシステムの出力とする．対話システムのコンテストであるローブナー賞で，2001 年に優勝した対話システム A.L.I.C.E. [Wallace 09] は，約 4 万件の用例を保持している．しかし，このような用例を人手で大量に作成するのはコストがかかる．

用例データの収集や作成について，既にいくつかの研究がなされている [Shawar 03, Huang 07]．これらの手法では，対話コーパスを利用することで大量の用例を収集できるものの，不特定多数のユーザから集めるため応答文の一貫性がなくなったり，過度に文脈に依存した応答文や不適切な応答文を出力する可能性がある．これらの問題を回避して用例を収集する枠組みの一つになりきり質問応答 [Higashinaka 13] がある．なりきり質問応答は，複数の作業者がキャラクタになりきって質問文に回答することで，作業者ごとの応答文の一貫性を担保する．また，複数の作業者が同時に一つのキャラクタの用例を作成することが可能になるため，一人当たりの作業量を分散させつつ，短時間で大量の用例を獲得することに成功している．しかしながら，この枠組みを用いても，収集できる用例には限りがあり，より多くの用例データを獲得するための手法が求められる．

本研究では，先行研究で収集された用例を元データとして，用例の応答文の一貫性を担保したまま，より大規模な用例へとデータを拡張する手法を提案する．この手法では，元データの用例の応答文に類似したリプライを持つツイート・リプライペアを検索し，そのツイートを応答文に対応する発話文として紐づけることで，応答文の一貫性を保ちながら用例を増やす．一般的な雑談対話システムが大量の発話文と応答文の対を元に，ある発話文に対して適切な応答文がどれであるかを推定する問題を解くが，本研究では，少量の応答文を元に，その応答文が利用可能な発話文のバリエーションを最大化する取り組みを行う．

2. 関連研究

2.1 用例ベース対話システム

先述の通り，雑談対話システムを実現する上で最も一般的な手法の一つに用例ベース対話システム [Murao 03] がある．用例ベース対話システムの持つ用例は，基本的には発話文 q と応答文 r の一対で構成され，用例 (q, r) は正しく対応しているものと仮定される．また，一つの発話 q に複数の応答 $r \in \mathbf{r}$ を紐付けた用例 $(q, \mathbf{r})$ を用いて，ユーザごとに最良と思われる応答文 r を出力毎に選び直す適応的応答選択という手法もある [Mizukami 15]．これらの用例ベース対話システムでは，ユーザから入力された発話文 q' に対して応答文 r が出力としての適切かを評価する関数 $f(r, q|q')$*1 を最大化する用例 (q, r) を探索することで，出力する応答文 r を決定する．

連絡先: 水上 雅博，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 ，mizukami.masahiro@lab.ntt.co.jp

*1 例えば，ユーザから入力された発話文 q' と用例の発話文 q とのコサイン類似度など

表 1: なりきり QA データの一例

なりきり対象	質問文 Q	応答文 A
A	ねぇ、どこ住み？てか LINE やってる？	通報しますよ？
	ねぇ、どこ住み？てか LINE やってる？	事務所を通してください
	見えてますよ？	ななななななな何がですか !?
M	ＬＩＮＥ使ってますか？	使ってるよ。LINE ブログもよろしく！
	いいたいことも言えない世の中は？	ポイズン！
	最後の晩餐は何食べる？	ラーメン！

表 2: なりきり QA データの詳細

キャラクタ：A	データ総数	13669
	ユニーク Q 数	5734
	ユニーク A 数	13146
	2 回以上登場した A の数	311
キャラクタ：M	データ総数	11639
	ユニーク Q 数	6376
	ユニーク A 数	10776
	2 回以上登場した A の数	400

2.2 用例の収集

用例ベース対話システムを構築する上で最も大きい課題の一つがデータ収集のコストである．既存の対話コーパスを利用することで大量の用例を収集する手法 [Shawar 03, Huang 07] も提案されているものの，応答文の一貫性が保証されず，不適切な応答を行う可能性があるという課題があった．これに対して，不適切な用例を事前に除外するための識別モデルを構築する手法 [水上 16] や，コーパス中の話者ごとに応答文生成のモデル学習することで応答文の話者に依存する一貫性を高める手法 [Li 16] が提案されている．また，雑談対話システムの応答を事前に規定された汎用的な相槌応答クラス（はい，そうだね，わかるよ，かっこいいね，など）に絞って応答選択を学習することで高い自然性を獲得する研究 [Inaba 16] もある．これらの手法は一般的な話者の発話文と応答文において，応答文の自然性，話者に依存する一貫性を高めることに成功しているものの，なりきり質問応答で得られる歴史上の人物や架空のキャラクタといった非常に強い特徴を持った応答文を得ることは困難である．

2.3 なりきり質問応答

用例データを人手で作る際にかかる大きなコストと、不特定多数のユーザから集められた用例に一貫性が担保されないという問題を解決するための枠組みの一つになりきり質問応答がある [Higashinaka 13]．なりきり質問応答は，作業者がキャラクタになりきって質問文に回答することで，質問文と「キャラクタらしい一貫性を持った応答文」を対にしたなりきり QA データを集めることができる．また，複数の作業者が同時に一つのキャラクタのデータを作成することが可能になるため，一人当たりの作業量を分散させつつ，短時間で大量の用例を獲得することに成功している．しかしながら，この枠組みを用いたとしても，収集できるなりきり QA データには限りがあり，既存の対話コーパスを利用することで大量の用例を収集する手法 [Shawar 03, Huang 07] に比べて獲得できる用例は少なくなる．用例の数が少ない場合，ユーザの入力に対して適切な応答文を出力することが困難になり，自然性が著しく低下する可能性がある．

3. なりきり QA データの収集

本研究の目的は，既知の用例を元データとして，用例の応答文の一貫性を担保したまま，より大規模な用例へとデータを拡張することである．本研究では，元となる用例として，なりきり QA のデータを利用する．

まず，2.3 節で説明した関連研究 [Higashinaka 13] と同様の枠組みで，複数の作業者からなりきり QA データを収集する．なりきりの対象となるキャラクタは 2 種類（A，M と呼ぶ）用意し，それぞれ A は 13669 件，M は 11639 件のなりきり QA データを収集した．表 1，2 にデータの詳細と一部例を示す．

4. 提案手法：用例の拡張

提案手法の具体的な処理内容について説明する．この手法は，元となる用例 $(q_{e,i}, r_{e,i}) \in \mathbf{e}$ の応答文 $r_{e,i}$ に対して，その応答文 $r_{e,i}$ が使えるような他の発話群 $q_{t,1}, q_{t,2}, ..., q_{t,n}$ を他の大規模な用例データ $(q_{t,i}, r_{t,i}) \in \mathbf{t}$ 中から探索し，新たに「応答文が使えそうな別の発話群 $q_{t,1}, q_{t,2}, ..., q_{t,n}$」と「元となる用例の応答文 $r_{e,j}$」をそれぞれ対として，新たに用例群 $(q_{t,1}, r_{e,i}), (q_{t,2}, r_{e,i}), ..., (q_{t,n}, r_{e,i})$ を作成することで元となる用例を拡張していく．応答文 $r_{e,i}$ と，その応答文 $r_{e,i}$ が使える発話文 $q_{t,j}$ が対応するかどうかは，元となる用例の応答文 $r_{e,i}$ と大規模な用例の応答文 $r_{t,j}$ との類似度 $sim(r_{e,i}, r_{t,j})$ に基づいて決定する．今回は，正規化レーベンシュタイン距離 [Yujian 07] が 0.1 以下であるという条件を設定した．この「元となる用例の応答文と，他の大規模な用例の発話文を対応させる処理」を，元となる用例 $(q_{e,i}, r_{e,i}) \in \mathbf{e}$ に対してそれぞれ行う．これによって得られた「元となる用例を拡張した用例の集合」を拡張用例と呼ぶ．

本論文では，元となる用例に 3. 節で説明したなりきり QA データを，大規模な用例データに Twitter から集めたツイート・リプライペアを利用した．Twitter ツイート・リプライペアはデータ総数 374M 件，ユニーク Q 数 334M 件，ユニーク A 数 289M 件のデータである．なお，ユーザを示すスクリーンネームを削除し，ハッシュタグおよび URL を含むツイートは除外している．図 1 に拡張手法の概要を示す．

次に，本手法による用例の拡張を行った結果を示す．なお，今回は処理の簡単化のために，利用する応答文をなりきり QA データの発話応答文のペア $(q_e, r_e) \in \mathbf{e}$ の応答文 r_e のうち 2 回以上登場したもののみとした．本手法による拡張の結果として，得られた拡張用例の一例を表 3 に，得られた拡張用例の詳細を表 4 に示す．

用例ベース対話システムの応答選択や，用例の獲得，高品質化の手法が発話文 q に対する応答文 r の妥当性を考える問題としているのに対して，提案手法は既知の応答文 r が利用できる発話文の範囲 $q \in \mathbf{q}$ を拡大するように用例を拡張する問題となっている．ここで注意したいのは，提案手法による用例

表 3: 拡張用例の一例

なりきり対象	質問文 Q	応答文 A
キャラクタ：A	サインください	事務所を通してください
	写メ撮った	事務所を通してください
	毎日 LINE するからな覚悟しとけよ	通報しますよ？
	オイ聞こえてんぞ	ななななな何がですか !?
キャラクタ：M	好きな食べ物は？	ラーメン！
	夜ご飯カレーとラーメンどっちがいいかな	ラーメン！
	山の日にも休めないこんな世の中じゃ	ポイズン
	真面目な人間ほど損をするこんな世の中じゃ	ポイズン

表 4: 拡張用例の詳細

キャラクタ：A	データ総数	1207k
	ユニーク Q 数	774k
	ユニーク A 数	131
キャラクタ：M	データ総数	9889k
	ユニーク Q 数	2102k
	ユニーク A 数	265

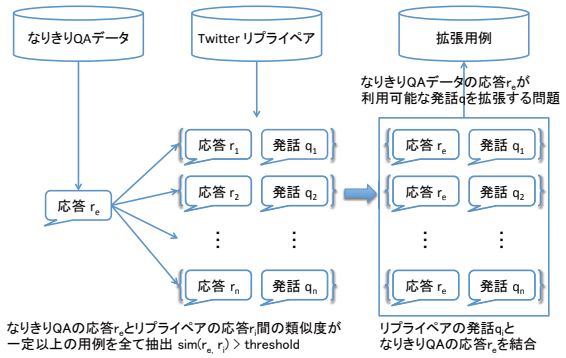

図 1: 用例の拡張手法

の拡張をどれだけ行っても，拡張用例に含まれる応答文 r は全て「元にした用例」に存在する応答文そのものであり，応答文の一貫性は常に保たれることである．

5. 実験

提案手法による効果を確認するために，4. 節で得られた拡張用例を対象に事例分析と定性評価，実際の対話システムへの組み込みによる主観評価を行う．

5.1 事例分析と定性評価

表 3 から，提案手法によってなりきり QA データの応答文と，その応答文が利用可能な発話文が対となった拡張用例が得られているのがわかる．表 3 と表 1 を比較すれば，「サインください」と「事務所を通してください」という拡張用例は，なりきり QA データの「ねぇ、どこ住み？てか LINE やってる？」と「事務所を通してください」という用例が元になっていることがわかる．また，拡張の際に利用されたツイート・リプライペアは「@(アカウント名) サインください」と「@(アカウント名) 事務所通してください」であった．これらの事例から，なりきり QA データの中の比較的汎用性の高い応答文が，なりきり QA データに存在しない発話文と組み合わせられ，新たな用例として拡張用例に含まれていることがわかる．

表 5: 主観評価結果

キャラクタ	手法	自然性	キャラクタ性
A	なりきり	3.04	3.15
	拡張用例	**3.23**	**3.24**
キャラクタ	手法	自然性	キャラクタ性
M	なりきり	3.16	3.17
	拡張用例	**3.39**	**3.20**

元となった用例の中でも汎用性の高い応答文と，多様な発話文が対応づけられることによって，より多様な場面でなりきり QA データの応答文が利用され，先行研究で課題となっていたユーザの入力に対して適切な応答文が見つからないケースを減少させる効果が期待される．

表 4 から，拡張用例は元にしたなりきり QA データに比べておよそ 100 倍程度のデータ数を持ち，ユニーク Q 数もキャラクタ A を対象にした際はおよそ 100 倍，キャラクタ M を対象にした際は 350 倍にも増加している．また，単語種類数においても，キャラクタ A を対象にした際はなりきり QA データの発話の単語種類が 6136 語であるのに対して，拡張用例では 153k 語，キャラクタ M を対象にした際はなりきり QA データの発話の単語種類数 6574 語に対して，拡張用例では 272k 語と増加している．

5.2 主観評価実験

主観評価実験では，2 つの対話システムを構築し，その出力について 26 名の被験者が評価を行った．主観評価の際には，用例の拡張の際に元にしたなりきり QA データの評価条件を踏まえ，テストデータ（なりきり QA データから抜き出した 50 個のデータであり，これは拡張用例および対話システムの学習時には用いられない）の発話文を入力として，出力文をそれぞれの対話システムから取得し，発話文に対する応答文の自然性とキャラクタ性について評価した．

実験に用いた対話システムは，全文検索とリランキングを用いた用例ベース対話システムである．まず，ユーザの入力文に対して最も近い発話文をもつ用例を全文検索を用いて取得する．この全文検索には Lucene[*2] を利用した．さらに用例の質問タイプや全文検索の一致度などを用いて，応答文をリランキングし，最も得点の高いものを出力する．2 つの対話システムの違いは利用している全文検索の対象およびリランキングの学習データで，なりきりではなりきり QA データのみを利用し，拡張用例ではなりきり QA データと拡張用例の両方を用いた．これらの対話システムによる主観評価結果を表 5 に示す．

表 5 より，キャラクタ A，M 共に拡張用例を用いた場合の

*2 https://lucene.apache.org/core/

方が高い自然性，キャラクタ性の評価を得た．特に，自然性の評価はなりきり QA データのみを用いた場合に比べて，拡張用例を用いた場合は有意に高くなった（Steel-Dwass 法による検定，$p < 0.05$）．これらの実験結果から，拡張用例はなりきり QA データのみを用いた場合と同程度のキャラクタ性を維持したまま，より自然な応答を可能にすることがわかった．

6. まとめ

本研究では，既知の用例を元データとして，用例の応答文の一貫性を担保したまま，より大規模な用例へとデータを拡張する手法を提案した．提案手法の有効性を示すため，なりきり QA データと Twitter ツイート・リプライペアを用いて実際に拡張用例を構築し，その性能を事例分析と定性評価，主観評価を通して検証した．実験結果から，用例の拡張は応答のキャラクタ性を維持したまま，自然性を向上させることが示され，これは提案手法の特徴と一致する．

提案手法の課題として，用例の拡張処理に用いる類似度とその閾値の設定に関する検証が挙げられる．本稿では，これらの設定をヒューリスティクスな仮定に基づき，類似度を正規化レーベンシュタイン距離および閾値を 0.1 以下と設定した．しかしながら，これらの設定は本来なら対話システムの性能を最大化するように決定されるべきである．そのため，対話システムの最終的な評価を最大化するように拡張用例の設定を決定する全体最適化の枠組みを考案していく必要がある．

謝辞

本研究では株式会社ドワンゴによる「なりきり質問応答」の企画で集まったなりきり QA データを利用しています．本データの収集にご協力いただきました株式会社ドワンゴ各位およびニコニコチャンネルユーザの皆様に感謝申し上げます．

参考文献

[Higashinaka 13] Higashinaka, R., Dohsaka, K., and Isozaki, H.: Using role play for collecting question-answer pairs for dialogue agents., in *INTERSPEECH*, pp. 1097–1100 (2013)

[Huang 07] Huang, J., Zhou, M., and Yang, D.: Extracting Chatbot Knowledge from Online Discussion Forums., in *IJCAI*, Vol. 7, pp. 423–428 (2007)

[Inaba 16] Inaba, M. and Takahashi, K.: Backchanneling via twitter data for conversational dialogue systems, in *International Conference on Speech and Computer*, pp. 148–155Springer (2016)

[Li 16] Li, J., Galley, M., Brockett, C., Spithourakis, P., Georgios, Gao, J., and Dolan, B.: A persona-based neural conversation model, *arXiv preprint arXiv:1603.06155* (2016)

[Mizukami 15] Mizukami, M., Kizuki, H., Nomura, T., Neubig, G., Yoshino, K., Sakti, S., Toda, T., and Nakamura, S.: Adaptive Selection from Multiple Reponse Candidates in Example-based Dialogue, in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (2015)

[Murao 03] Murao, H., Kawaguchi, N., Matsubara, S., Yamaguchi, Y., and Inagaki, Y.: Example-based spoken dialogue system using WOZ system log, in *Proceedings of the Fourth SIGdial Workshop of Discourse and Dialogue* (2003)

[Ritter 11] Ritter, A., Cherry, C., and Dolan, W. B.: Data-driven response generation in social media, in *Proceedings of the conference on empirical methods in natural language processing*, pp. 583–593Association for Computational Linguistics (2011)

[Shawar 03] Shawar, A., Bayan and Atwell, E.: Using dialogue corpora to train a chatbot, in *Proceedings of the Corpus Linguistics 2003 conference*, pp. 681–690 (2003)

[Wallace 09] Wallace, S., Richard: The anatomy of ALICE, in *Parsing the Turing Test*, pp. 181–210, Springer (2009)

[Wong 12] Wong, W., Cavedon, L., Thangarajah, J., and Padgham, L.: Strategies for mixed-initiative conversation management using question-answer pairs, *Proceedings of COLING 2012*, pp. 2821–2834 (2012)

[Yujian 07] Yujian, L. and Bo, L.: A normalized Levenshtein distance metric, *IEEE transactions on pattern analysis and machine intelligence*, Vol. 29, No. 6, pp. 1091–1095 (2007)

[水上 16] 水上 雅博, Graham, N., 吉野 幸一郎, Sakriani, S., 鈴木 優, 中村 哲：快適度推定に基づく用例ベース対話システム, 言語処理学会第 22 回年次大会 (2016)

[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)