対話システムのための「なりきり質問応答」を用いた質問応答ペアの収集とその応用

東中竜一郎*　　堂坂浩二　　磯崎秀樹

日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
higashinaka.ryuichiro@lab.ntt.co.jp, {dohsaka,isozaki}@cslab.kecl.ntt.co.jp

1　はじめに

雑談のような，制約の少ない対話を行うシステムは，扱う内容が多様なために，統一的な手法ではユーザ発話に適切に応答できないことがある．そのため，システムは，入力であるユーザ発話と，ユーザ発話対してシステムが出力すべき応答発話のペア（発話ペア）をあらかじめ大量に作成しておき，応答に利用することが多い．

例えば，対話システムのコンテストであるレブナー賞で，2001 年に優勝したシステム A.L.I.C.E. [7] は，発話ペアを約 4 万件保持している．また，音声対話システムの「たけまる君」も内部に発話ペアを多く備えている [5]．しかし，このようなデータを人手で大量に作成するのはコストがかかる．

発話ペアの自動作成については，既にいくつかの手法が提案されている [4, 2]．しかしながら，これらの従来手法は，不特定多数のユーザの対話データから自動的に発話ペアを抽出するため，ノイズが多く，また，応答に一貫性がないという問題がある．このようなデータに基づいて対話システムを構築すると，ユーザの入力に対し，不適切な応答や，矛盾した応答を行う可能性がある．

本稿では,「なりきり質問応答」という枠組みにより，特定の人物に紐付けられた発話ペアを効率的に収集する手法を提案する．具体的には，ユーザに「なりきり」によるやり取りを質問応答の形式で行ってもらうことを趣旨としたウェブサイトを構築し，発話ペアとして，質問応答ペアを収集する．

本手法は，なりきりの特性である好きな人物になりきる楽しさと，好きな人物に対して質問できるという面白さから，多数のユーザの参加を促し，多くの質問応答が効率的に収集できる可能性がある．また，複数のユーザが一人の人物になりきって質問に答えることで，回答者の負荷の軽減ができ，その結果，応答の迅速性が高まることから，効率的な質問応答ペアの収集が可能になると考えられる．近年，ユーザ参加型のデータ収集には，面白さが必要との指摘もあり [6]，本手法はこの流れに沿ったものと考えることができる．

質問対象: マリーアントワネット (user25)
Q1: ふりかえってみて貴女の一生はどうでしたか。(user25)
1. 貴方たちが思うほど不幸でもなかったわ (user25)
2. 幸福でした。子供たちに感謝してるの (user3)
3. とても楽しかったわ。(user13)
Q2: 結婚相手を選べるとしたら、誰と結婚したいですか (user19)
1. やっぱりフェルゼンかしら。あの方は私の命でした。夫には申し訳ないけど。(user21)
2. 結婚相手は誰でもよかった、楽しく毎日送れるならね。(user25)

図 1: マリーアントワネットについてのなりきり質問応答

図 1 は,「なりきり質問応答」によって，マリーアントワネットについて得られたなりきり質問応答の例である．括弧内は人物，質問，または，回答を投稿したユーザ ID を示す．複数のユーザが協力して一人の人物の質問応答ペアを作成している様子が分かる．

ウェブ上の掲示板の中には「なりきり」を趣旨としたものも存在する．しかしながら，自由に会話をしてもよいというサイトの性質から，再利用可能なの発話ペアの獲得には利用しにくい．また，専門家がユーザの質問への回答を掲載しているウェブサイトも多く存在する．しかしながら，一人の人間が答えられる質問の数には限界があり，対話システムに必要な分量の質問応答ペアを集めるにはコストと時間を要する．

2　データ収集実験

ユーザに，なりきりによるやり取りを質問応答の形式で行ってもらう趣旨のウェブサイト「なりきり質問応答」を構築し，トライアル実験として，50 人の

*現在，NTT サイバースペース研究所

登録された人物数	397
質問数	2,502
質問数（重複なし）	2,483
回答が一つ以上寄せられた質問数	2,262
一人物に寄せられた平均質問数	6.42
回答数	3,838
自演応答数	213
一質問あたりの回答数	1.70
一人物あたりの回答数	9.99
投稿（質問・回答）の総数	6,340
一質問に回答したユーザ数	1.62
一人物あたりの回答ユーザ数	5.21
一ユーザあたりの質問数	49.80
一ユーザあたりの回答数	76.04
一ユーザあたりの投稿数	125.84
150 以上の投稿をしたユーザ数	9
200 以上の投稿をしたユーザ数	3

表 1: 投稿された質問応答データの統計情報

ユーザを募り，実際にサイトを 3 週間にわたって使用してもらった．ここでは，そのトライアル実験について説明する．なお，実験に際してはサイトにパスワードをかけ，一般には非公開で行った．

トライアル実験は，例示として，4 人物，12 の質問，36 の回答が寄せられたサイトの状態からスタートさせた．実験に参加してもらう各ユーザには謝礼を支払い，期間中，最低 100 の投稿（質問，または，回答）をするように，また，質問と回答の数があまり偏らないように教示した．ユーザは自由に人物を登録できる．なお，自演応答（自分の質問に自分で回答すること）は投稿数にカウントしないこととした．

3 週間のトライアルで収集された質問応答データの統計情報は表 1 の通りであった．50 人のユーザがそれぞれ 100 の投稿を行うと，5000 の投稿が集まることになるが，期間内に，6340 の投稿が集まった．これは，なりきりによる質問応答のやり取りにユーザが楽しみを見出し，積極的に質問応答に参加したことによると考えられる．例えば，150 以上の投稿を行ったユーザは 9 人，200 以上の投稿を行ったユーザは 3 人いた．また，投稿数にカウントされないにも拘わらず，自演応答の投稿数が 213 あった．

一人物の回答に関わったユーザ数は 5.21 であり，一つの質問は平均 1.62 ユーザが回答している．このことから，複数人が一人以上のユーザが同一人物になりきって質問に回答していることが分かる．一人だけが回答し続けるのと異なり，複数のユーザにより，回答の負荷を分散しつつ質問応答がなされていることが分かる．

表 2 は，トライアル実験後に行ったユーザアンケートの結果である．このアンケートは 5 段階評価であり，5 が一番良い．Q1–Q3 の結果から，ユーザはなりきり質問応答サイトを楽しんだことが分かる．謝礼を払っているとはいえ，5 段階評価で 4.28 という

	設問	平均 (標準偏差)
Q1	サイトの使いやすさ	3.90 (0.92)
Q2	サイトをまた使ってみたいか	3.98 (1.12)
Q3	なりきりを楽しめたか	4.28 (0.72)

表 2: トライアルサイトのアンケート結果

高い平均点を得ていることは，なりきり質問応答がユーザを引きつけ，質問応答データを能動的に提供するように強く動機づけることができたと言える．

これらの統計情報やユーザアンケートの結果から，なりきり質問応答は人物に紐づけられた質問応答ペアの収集に有効であると考えられる．

3 収集した質問応答ペアの応用

対話システムは，A.L.I.C.E. やたけまる君がそうであるように，個々の個性に紐づけられた発話ペアを保持するが，システムの発話ペアにも網羅性に限界があり，必ずしもユーザのすべての入力に対応できるわけではない．ある対話システムに対してなされた質問が，そのシステムの発話ペアに見つからず，応答できないこともある．そのような場合に，他のシステムが適切な発話ペアを保持していれば，それを代用して応答することで，システムの応答能力が向上する可能性がある．

ここでは，本手法で集めた質問応答ペアを用い，ある人物が未知の質問をされたとき，他の人物の質問応答ペアを用いてその質問に答えることができるかを実験によって確かめることで，システム間の発話ペアの相互利用の可能性を検証する．今回，他の人物の質問応答ペアを用いる応答手法として，下記の 4 手法を考えた．

単語類似度: 人物 P に対する未知の質問 Q について，人物 P 以外の人物になされた質問から，Q に最も単語が類似した質問 Q' を検索し，Q' に対する回答から回答を一つ選択し，回答とする．ここで，Q と Q' の類似度は，下記の式により算出される，それぞれの質問に含まれる単語のセットのコサイン類似度を用いる．words とは質問中の内容語のセットをバイナリのベクトルとして返す関数である．

$$\mathrm{sim}(Q, Q') = \frac{\mathrm{words}(Q) \cdot \mathrm{words}(Q')}{\mid \mathrm{words}(Q) \mid\mid \mathrm{words}(Q') \mid}$$

Q' が見つかったとして，Q' に対する回答が複数ある場合がある．この中から一つを選択するため，Q' がなされた人物と最も関連が低いものを一つ選ぶ．これは，回答には質問された人物と関連が深い単語が入ることが多く，そのような単語が入った回答は P の回答として不適切だと考えられるからである．この回答選択のプロセスは下記のように書ける．

$$\mathrm{argmin}_{a \in A} \ \ \mathrm{maxpmi}(\mathrm{person}(a), a)$$

ここで，A は Q' に対する複数の回答セットを指し，a はその要素である．maxpmi は人物名と回答との関連の強さを Pointwise Mutual Information (PMI) により計算し，その最大値を返す関数であり，person は回答に紐付けられた人物名を返す関数である．なお，PMI は，人物名 ($\mathrm{person}(a)$) と回答中の内容語 ($w_i \in a$) との間で，下記の式により計算する．

$$log_2(\frac{\mathrm{docs}(\{\mathrm{person}(a), w_i\})/N}{\mathrm{docs}(\mathrm{person}(a))/N \cdot \mathrm{docs}(w_i)/N})$$

ここで，docs は特定のテキスト文書中における，引数の人名または単語を含む文書数である．N はテキスト文書中の総文書数である．本実験ではテキスト文書群としてウィキペディアを用いた．なお，類似した質問を検索し，その回答を用いて応答するシステムとして FAQFinder [1] があるが，本手法では回答する人物も考慮し回答を選択している点が異なる．

質問分類制約: 単語類似度と同様，コサイン類似度を用いた質問の検索をするが，その際に，質問分類が同じもののみを検索対象とする．ここでは，紙面の制約により詳細は割愛するが，質問分類が同じとは，質問の意図や質問に含まれる人物属性が同じであることを指す．後述する評価実験では，手動でラベル付けした質問分類を用いるが，ある程度の分量のデータがあれば，自動での質問分類のラベル付けは可能である．

人物間類似度: 質問の検索に単語類似度だけではなく人物間の類似度も利用する．例えば，織田信長に対する質問と類似した質問を探す際には，俳優やタレントへの質問からではなく，明智光秀や石田光成といった武将に対する質問を重視して検索する．ここで，質問の類似度は下記の式により算出する．

$$\mathrm{sim}(Q, Q') - \mathrm{dist}(\mathrm{person}(Q), \mathrm{person}(Q'))$$

ここで，dist とは人物間の距離を返す関数で，テキストデータ（ここでは，ウィキペディア）における人物名の共起から計算される．dist は Q と Q' が近ければ近いほど小さくなる．具体的には，まず，松林らの手法 [3] に従い，フィッシャーの正確確率検定により得られる人名同士の共起を，2 次元上のグラフとして表現する．dist はこのグラフ上の人名間のユークリッド距離の自然対数を返す．

質問分類制約＋人物間距離: 質問を検索する際に，質問分類による制約を満たすものの中で，人物間類似度の手法により最も類似する質問を検索する．

手法	内容妥当性	表現妥当性
オリジナル回答	5.33	5.50
単語類似度	2.46	3.39
質問分類制約	2.67	3.55
人物間距離	2.34	3.75
質問分類制約＋人物間距離	2.71	3.77
オラクル	3.44	4.62

表 3: 各手法の回答に与えられたスコア

3.1 評価実験

上記 4 手法の性能を検証するため，人間の評価者による評価実験を行った．トライアル実験で得られた全質問の内，実際にユーザによって回答が一つ以上なされ，かつ，4 手法のすべてが何らかの応答を返すことができた 967 の質問に対して，各手法が出力した回答を 3 人の評価者が独立に評価した．

評価者は質問とその質問がなされた人物を提示され，4 手法が出力した回答のそれぞれについて，内容妥当性（内容がその人らしいかどうか），および，表現妥当性（表現がその人らしいかどうか）という観点から，7 段階で評価した．また，4 手法に加え，対象の質問に対し実際にトライアル実験のユーザが行った回答（オリジナル回答）のうち一つについても評価してもらった．オリジナル回答が複数ある場合，対象の人物名と最も PMI の高い単語を含む回答を選んだ．

表 3 に 3 人の評価者による平均スコアを示す．オリジナル回答のスコアの高さから，トライアル実験でのユーザの回答が適切なものであったことが分かる．4 手法については改善の余地があることが分かるが，統計的検定（Steel-Dwass の多重比較）を行った結果，質問分類による制約を用いると内容妥当性が有意に改善し ($p<0.01$)，人物間距離を導入すると，表現妥当性が有意に改善することが分かった ($p<0.05$)．ここで，オラクルとは，各質問について，4 手法から最もよい回答を選択できた場合のスコアを指す．オラクルのスコアは比較的悪くないことから，質問によって高精度に手法が選択できれば，比較的適切な回答が自動的に行えることが分かる．

図 2 は，提案 4 手法が出力した応答の例である．全く異なる人物への質問の回答を用いても，応答として適切なものを出力できることが分かる．なお，スコアは評価者の主観に大きく依存する．例えば，質問分類制約＋人物間距離の内容妥当性に着目すると，3 人の評価者の内容妥当性スコアはそれぞれ，3.99, 2.28, 1.85 と大きく異なる．

質問を検索する際にほとんど単語が重複していないものから探しても見当違いの回答をしてしまう恐れがある．また，人物間距離があまりに大きい人物への質問を用いても，対象人物らしい表現を持つ回

質問対象：ビル・クリントン
質問：「影武者はいますか？」

オリジナル回答: それは答えられないな。

単語類似度: 危なくなったときに、かくまってくれそうな人なら何人か。(「杉田かおる」の「仲のいい友達はいますか？」への回答)

質問分類制約: えー作るヒマないですよぉー (「安藤美姫」の「彼氏はいますか？」への回答)

人物間類似度: 決してゆだんできぬヤツだな！(「チャーチル」の「ルーズベルトをどう思いますか?」への回答)

質問分類制約＋人物間距離: 敵が多すぎて特定は出来無い。尊敬するリンカーンの末席に並べたのが幸せ。(「ジョン・F・ケネディ」の「あなたを撃った人間に心当たりはありますか…？」への回答)

図 2: 提案 4 手法を用いた質問応答の例

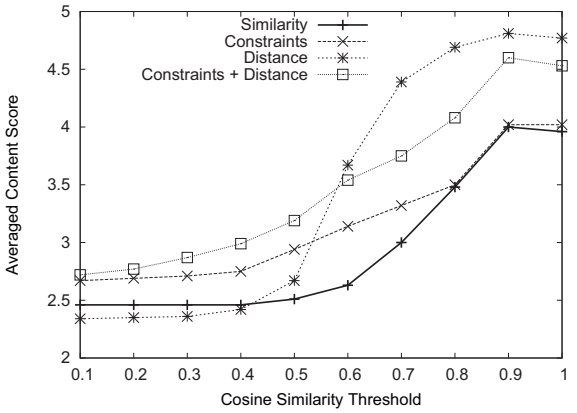

図 3: コサイン類似度の閾値 (Cosine Similarity Threshold) を変動させた場合の内容妥当性 (Averaged Content Score) の推移．Similarity，Constraints，Distance，Constraints+Distance はそれぞれ，単語類似度，質問分類制約，人物間距離，質問分類制約＋人物間距離を指す．

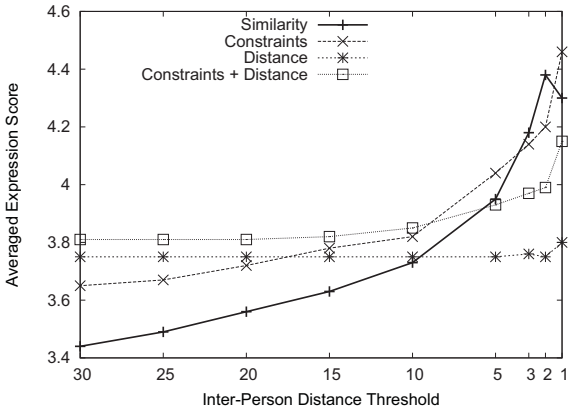

図 4: 人物間類似度の閾値 (Inter-Person Distance Threshold) を変動させた場合の表現妥当性 (Averaged Expression Score) の推移．

答が得られない可能性が高い．そこで，質問類似度と人物間距離に閾値を設定し，閾値以上の質問からしか検索しないようにした場合に，内容妥当性や表現妥当性のスコアがどう変化するかを調べた．図 3 および図 4 は，コサイン類似度と人物間距離の閾値を変動させた場合の内容妥当性および表現妥当性の推移のグラフである．

これらのグラフから，コサイン類似度の閾値を 0.6 程度に設定することで，内容妥当性については，問題のないレベルと考えられる 4 に近いスコアを実現できることが分かる．また，人物間距離の閾値を 3 程度に設定することで，こちらも 4 程度の表現妥当性を実現できることも分かる．

もちろん，閾値を変動させると回答可能な質問数が減少する．例えば，閾値が 0.6 の場合，質問分類制約＋人物間距離は 967 の質問の内，105 のみにしか回答が出力できない．しかし，全体の 10%であっても，質問応答ペアの相互利用が可能であるという実験結果は，今後の対話システムの応答性能の向上に繋がるものである．

4 まとめと今後の課題

本稿では「なりきり質問応答」を用いた質問応答ペアの収集と，その応用として複数の人物に紐付けられた質問応答ペアの相互利用について述べた．今後は，データ収集をより大規模で行うとともに，相互利用のための手法の改善を行っていく予定である．

参考文献

[1] R. D. Burke, K. J. Hammond, V. A. Kulyukin, N. Tomuro, and S. Schoenberg. Question answering from frequently asked question files: Experiences with the FAQFinder system. Technical report, University of Chicago, 1997.

[2] J. Huang, M. Zhou, and D. Yang. Extracting chatbot knowledge from online discussion forums. In *Proc. IJCAI*, pp. 423–428, 2007.

[3] T. Matsubayashi and T. Yamada. A force-directed graph drawing based on the hierarchical individual timestep method. *International Journal of Electronics, Circuits and Systems*, 1(2):116–121, 2007.

[4] B. A. Shawar and E. Atwell. Using dialogue corpora to train a chatbot. In *Proc. International Conference on Corpus Linguistics*, pp. 681–690, 2003.

[5] S. Takeuchi, T. Cincarek, H. Kawanami, H. Saruwatari, and K. Shikano. Construction and optimization of a question and answer database for a real-environment speech-oriented guidance system. In *Proc. COCOSDA*, 2007.

[6] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proc. CHI*, pp. 319–326, 2004.

[7] R. S. Wallace. *The Anatomy of A.L.I.C.E.* A.L.I.C.E. Artificial Intelligence Foundation, Inc., 2004.