1H2-7

Q&Aコミュニティでの質疑応答パターンの理解

Understanding of Writing Style Patterns between Q&A in Knowledge Sharing Community

西原陽子[*1] 松村真宏[*2] 谷内田正彦[*3]
Yoko NISHIHARA Naohiro MATSUMURA Masahiko YACHIDA

[*1]東京大学大学院工学系研究科
School of Engineering, The University of Tokyo

[*2]大阪大学大学院経済学研究科
Graduate School of Economics, Osaka University

[*3]大阪工業大学情報科学部
Faculty of Information Science and Technology, Osaka Institute of Technology

Knowledge sharing communities, which are supported by Internet users, are on the Web. In the knowledge sharing communities, categories denoting question themes are prepared. The Internet users have to select a category to send questions. Each category has its own different group of the users. It is considered that each category has its own user's writing style patterns because people change their speech sytles depending on their hearers.

In this paper, we surveyed the writing style patterns between questions and answers sent to a Knowledge sharing community, Yahoo! Chiebukuro. From our analysis, following four writing style patterns are presented: (1) if a user questions considerately, the user will be answered considerately, (2) if a user questions to require ideas, the user will be answered cooperative ideas, (3) if a user questions to require a right answer only, the user will be answered the right answer only, and (4) if a user questions to require ideas, the user will be answered ideas from different viewpoints.

1. はじめに

Web 上に，質問を投稿すると他の人が回答を投稿してくれる知識共有コミュニティがある．知識共有コミュニティには，質問の主題を表す複数のカテゴリが用意されており，ユーザは質問の主題に適したカテゴリを選んでから投稿することになっている．質問や回答を投稿するユーザはカテゴリごとに異なり，人間は聞き手に応じて発言の文末表現を変えることから，カテゴリごとに質問と回答の文末表現も異なると考えられる．

そこで本稿では知識共有コミュニティの一つである Yahoo! 知恵袋[*1]に投稿された質問と回答に存在する文末表現のパターンをカテゴリごとに調査した結果を報告する．本稿では質問者が気に入って選んだ回答をベストアンサーと定義し，文末表現のパターンを「質問とベストアンサーにのみ含まれる文末表現」と定義する．調査では，質問と回答から文末表現を抽出し，質問とベストアンサーのペアをクラスタリングする．得られたクラスタの中で，SVM を用いてベストアンサーとそれ以外の回答を学習し，分類精度を評価する．同時にクラスタリングしない場合の分類精度も評価し，クラスタリングした場合の分類精度が高ければ，そのクラスタに文末表現のパターンが存在すると見なして，文末表現を抽出する．

知識共有コミュニティでは書き言葉を用いてコミュニケーションが行われている．書き言葉によるコミュニケーションを分析する研究では，発話タグを用いることが多い [Core97, Jurafsky97]．分析の実例には，電子メールを用いて行った２つの国際会議の統合が成功した原因を調査した研究や [山下 02]，会話の盛り上がりの原因を分析した研究 [徳久 06] などがある．これらの研究では発話タグを人手で与える必要があり，多くの時間を要すると予想される．本稿では文末表現のパターンを抽出する作業は計算機によって行い，抽出された文末表現のパターンを分析する作業は人手によって行う．一部の作業を計算機に任せることで，多くのカテゴリの比較，調査を実現する．

国際会議 TREC や国内ワークショップ NTCIR では，質問に最適の回答を抽出するための手法が研究されている．質問のタイプを 5W1H で分類し，検索範囲を狭めた上で回答を抽出する手法 [Hovy00] や質問から特定の文や段落といったパッセージを抽出し，パッセージを検索に用いることで回答を抽出する手法が提案されてきた [Hearst93]．これらの手法では質問と回答で類似するキーワード（名詞，動詞，形容詞）に注目することが多かった．これに対し，本稿ではキーワードとして抽出されることの少ない文末表現を抽出し，質問と質問者が気に入る回答にのみ存在する文末表現のパターンを調査する．

連絡先: 西原陽子，東京大学大学院工学系研究科，〒 113-8656
東京都文京区本郷 7-3-1，nishihara@sys.t.u-tokyo.ac.jp

*1 http://chiebukuro.yahoo.co.jp/

2. 文末表現のパターン抽出方法

文末表現のパターンを抽出する方法を説明する．初めに質問と回答から文末表現を抽出する．次に，抽出された文末表現を属性とし，質問とベストアンサーのペアを表すベクトルを作り，ベクトルをクラスタリングする．得られたクラスタの中で，SVM を用いてベストアンサーとそれ以外の回答を学習し，分類精度を求める．同時にクラスタリングしない場合の分類精度も評価し，クラスタリングした場合の分類精度が高ければ，そのクラスタには文末表現のパターンが存在すると見なし，クラスタに特徴的な文末表現を抽出する．

2.1 質問と回答の例

質問と回答の例を表 1 に示す．本稿では質問者によって内容が比較，判断された回答に限定し，文末表現のパターンを抽出したいため，用いる質問と回答には，**(1) １つの質問には少なくとも異なる２人からの回答がある**，**(2) ベストアンサーは最初または最後に投稿されたものではない**，と２つの条件を設ける．

表 1: 質問 (Q)，ベストアンサー (BA)，ベストアンサー以外の回答 (NA) の例

Q	ゾウが描いた絵があると聞きました．本当ですか？あるのなら何処で見れるのでしょうか？
BA	少し前に「笑っていいとも」に高橋克実さんがゲストで出演なさった際に，紹介されました．(星になった少年の PR でした) ゾウが描いたとは思えない絵でしたよ．人間が筆を渡したら，自分で勝手に描くそうです．普通の動物園のゾウが描いた絵でしたら,「ゾウ描いた絵」等のキーワードでヒットすると思います．こちらは，鼻で直接描いた抽象画のようなものが多いです．
NA	あるといえばあるけど，絵というよりは筆を叩きつけてるだけみたいですよ？ニュースとかで見れると思います...

2.2 文末表現の抽出

質問と回答から文末表現を抽出する．本稿では文末表現を「文の文末から文頭に遡って出現する助詞，助動詞の連なり」と定義する．文末表現は質問と回答の中の全ての文から茶筌 [松本 97] を用いて抽出する．文数の少ない質問，回答からも多くの文末表現を得たいため，1,2,3-gram の全てを抽出する．例えば表 1 の質問 Q からは「た，ます，ますた，か，です，ですか，う，うか，ですう，ですうか」が抽出される．質問と回答から抽出された文末表現は区別し，質問から抽出された場合には「Q です」，回答から抽出された場合には「A です」と表記する．文末表現は種類が多くなるため，抽出された合計の多い順に上から全体の T_1 %[*2]に相当するものだけを用いる．

2.3 質問と回答のクラスタリング

抽出された文末表現を用いて，質問と回答をクラスタリングする．クラスタリングするために，質問とベストアンサーをベクトル QBA で表現する．

$$QBA(q_i, ba_i) = (x_1, \ldots, x_M, x_{M+1}, \ldots, x_{M+N}) \quad (1)$$

式 (1) では，q_i が質問，ba_i はベストアンサー，x_1 から x_M は全ての質問から抽出された文末表現，x_{M+1} から x_{M+N} は全ての回答から抽出された文末表現とし，ベクトルの次元は $M+N$ 次元とする．x の値は q_i と ba_i 中の頻度とする[*3]

次に，ベクトル QBA を階層的にクラスタリングする．2 つのクラスタの距離は Ward 法で測る．Ward 法によるクラスタ C, C' 間の距離 d は式 (2) によって表される．

$$d(C, C') = E(C \cup C') - E(C) - E(C') \quad (2)$$

式 (2) では $E(C)$ がクラスタ C 中の全てのベクトルから C の重心までの距離の二乗の総和を表す．2 つのクラスタを 1 つにするときの距離には閾値を設け，閾値 T_2[*4]を超えた場合にクラスタリングを終了する．

2.4 文末表現のパターンの抽出

質問とベストアンサーにのみ存在する文末表現のパターンを抽出する．文末表現のパターンを抽出するために，クラスタごとにベストアンサーとそれ以外の回答を SVM[Vapnik95] によって機械学習し，分類精度を測る[*5]．質問と質問者が選ばなかった回答をベクトル QNA で表現する．

$$QNA(q_i, na_i) = (x_1, \ldots, x_M, x_{M+1}, \ldots, x_{M+N}) \quad (3)$$

*2 3 章の分析では $T_1 = 10$ とした．
*3 x の値には TF-IDF 値や IDF 値も考えられるが，手法の単純さから本稿では頻度を用いる．
*4 3 章の分析では $T_2 = 100$ とした．これは著者らが目で見て，これ以上まとめることに意味がないと思った値になる．
*5 フリーソフトウェア SVM^{light} の Inductive SVM を用いた．

表 2: 本稿で用いた Yahoo!知恵袋の質問のカテゴリ

子育て、出産	芸術、文学、歴史
病気、症状、ヘルスケア	健康、病気、ダイエット
不動産、引っ越し	ファッション
交通、地図	恋愛相談、人間関係の悩み
海外	結婚
妊娠、出産	ゲーム
新車	自動車
家事	バイク
国内	料理、グルメ、レシピ
動物、植物、ペット	レシピ、調理法
Yahoo!オークション	コミック
小・中学校、高校	一般教養
Yahoo!知恵袋	子育ての悩み
政治、社会問題	マナー
パソコン、周辺機器	芸能人、タレント
野球	テレビ、ラジオ
インターネット	話題の人物
言葉、文学	

表 3: クラスタリングに用いた文末表現

質問	Q う	Q うか	Q か	Q ますた
	Q くらい	Q た	Q たい	Q ますん
	Q だ	Q です	Q ですう	Q まで
	Q ですうか	Q ですか	Q ない	Q ますか
	Q など	Q ます		
回答	A うか	A か	A くらい	A ますん
	A た	A たい	A だ	A まで
	A だけ	A だない	A です	A ますた
	A ですう	A ですうか	A ですた	A ます
	A ない	A ないです	A など	A ぬ

分類精度を比較するため，得られたクラスタ内の質問と回答の数を揃えた新しい集合を用意し，同様に分類精度を評価する．クラスタリングせずに集められた質問と回答の集合では，特徴的な文末表現のパターンは存在しないと考えられる．そこで，クラスタリングした場合の分類精度が高ければ，文末表現のパターンが存在すると見なして，抽出する．クラスタリングした場合としない場合での分類精度を比較する際，有意差はカイ 2 乗検定を用いて評価する．

最後にクラスタリングした場合の分類精度が高かったクラスタから，文末表現のパターンを抽出する．ここでは，クラスタの重心を表すベクトルから，それぞれの属性の値を取り出し，取り出した値を他のクラスタの同じ属性の値と比較し，取り出した値が有意に高ければ，その属性を文末表現のパターンとして抽出する．有意差はカイ 2 乗検定を用いて評価する．

3. 文末表現のパターンの分析

2 章で説明した方法を用いて文末表現のパターンを抽出し，分析を行った．分析では，Yahoo!知恵袋に 2005 年 9 月度に投稿された質問，回答を用いた．本稿では，表 2 に示す質問数が十分にある 35 個のカテゴリに限定し，1 つのカテゴリから 2.1 の条件を満たす質問と回答を 1,000 件ずつ，合計 $3,000 \times 35 = 105,000$ 件用意した．クラスタリングには表 3 に示す 38 種類の文末表現を用いた．

3.1 抽出された文末表現のパターン

クラスタリングする場合としない場合での，SVM によってベストアンサーとそれ以外の回答を分類する精度の平均を図 1 に示す．図 1 では，クラスタごとに得られた分類精度の平均を

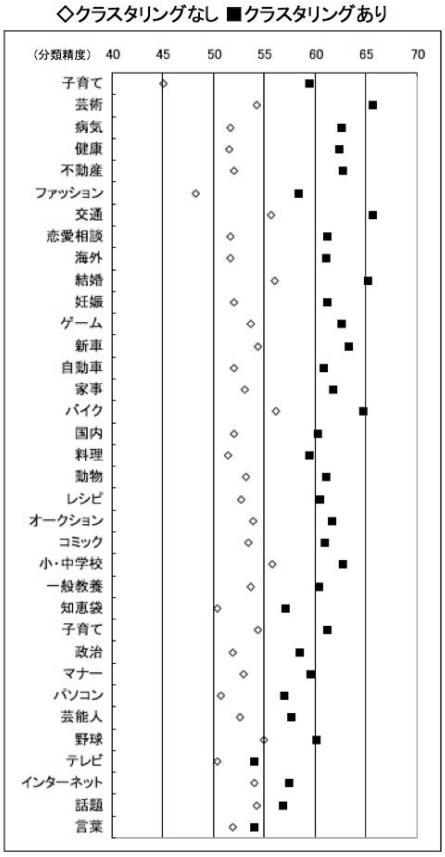

図 1: ベストアンサーとそれ以外の回答の分類精度

とった．図 1 では全てのカテゴリでクラスタリングした方の分類精度が高かった (P<.05)．この理由は，質問とベストアンサーにのみ存在する文末表現のパターンがあったためと考えられる．このことから，質問とベストアンサーにのみ存在する文末表現のパターンがあると分かった．

各クラスタから抽出された文末表現のパターンを表 4 に示す．クラスタごとに得られた特徴的な文末表現はカテゴリごとにほぼ同じものが得られたため，表 4 には全てのカテゴリに共通したものを示している．クラスタリングの結果，1 つのカテゴリから 3 個，または 4 個のクラスタが得られた．5W1H では，質問のタイプを 6 種類に分類できるが，文末表現で分類すると質問は 4 種類に分類できることが分かった．

3.2 文末表現のパターンの解釈

クラスタ C1 については，各カテゴリから 30 件ずつ質問と対応する回答を取り出し，検証したところ，平均 25 件の質問で「です，ます」を用いて，丁寧な表現で記述されており，ベストアンサーでも「です，ます」を用いて丁寧な表現で記述されていた．すなわち，**質問者の丁寧な記述に対し，回答者も丁寧な記述で返す**パターンと解釈できる．

クラスタ C2 については，30 件中，平均 21 件の質問で質問者はある事柄に対する自分の意見や感想を記述し，他の回答者の意見や感想を問う記述をしており，ベストアンサーでも回答者は質問者に協調する意見や感想を記述していた．すなわち，**質問者が意見や感想を記述することに対し，回答者は協調する意見や感想を返す**パターンと解釈できる．

クラスタ C3 については，30 件中，平均 18 件の質問で文末表現が用いられず，簡潔に方法や事物について質問されており，ベストアンサーでも同様に簡潔に質問者の質問に回答がされていた．すなわち，**質問者が答を問うことに対し，回答者が答のみを返す**パターンと解釈できる．

表 4: 得られた文末表現のパターン

クラスタ	文末表現のパターン（上段）と解釈（下段）
C1	Q です，Q ます，A です，A ます，A か，が多い． **質問者の丁寧な記述に対し，回答者も丁寧な記述で返す．**
C2	Q ますか，A ないです，A ました，が多い． **質問者が意見や感想を記述することに対し，回答者は協調する意見や感想を返す．**
C3	（なし） **質問者が答を問うことに対し，回答者が答のみを返す．**
C4	Q だ，Q たい，Q ですか，Q ない，A だ，A ない，A ですか，が多い． **質問者が意見，感想を記述することに対し，回答者は異なる観点からの意見や感想を返す．**

クラスタ C4 については，30 件中，平均 20 件の質問では，質問者がある事柄に対する自分の意見や感想を述べ，他の回答者の意見や感想を問う記述になっており，ベストアンサーでは，質問者とは異なる観点からの意見や感想が記述されていた．すなわち，**質問者が意見，感想を記述することに対し，回答者は異なる観点からの意見や感想を返す**パターンと解釈できる．

得られたパターンと従来の知見を比較すると，クラスタ C1，C2，C3 では，質問者と回答者の文を記述する態度が似ており，これは「態度の類似性」[Byrne71] に一致する．「態度の類似性」は人間関係の中で，類似する態度や意見が多いほど，その人に対する好意も増加するという知見だが，質問者が回答を選ぶ状況では，複数の回答者がいる中で類似する態度を示した回答者に対する好意が大きくなり，その回答者の回答を気に入って選んだと解釈できる．従来の知見に沿うことから，本稿で得られた文末表現のパターンは妥当なものと考えられる．

クラスタ C4 では，「態度の類似性」には一致しないが，質問者と複数の回答者で問題解決を行う中で，自分の意見や感想と異なる観点からの意見や感想が出された時に，質問者はその意見や感想を自分にとっても有益な情報になると評価する傾向があったと解釈できる．

4. 考察

クラスタリングでは質問，回答が各クラスタに均等に分類されたわけではなく，その数には偏りがあった．1 個のクラスタに 400 件以上の質問が含まれたカテゴリを表 5 に示す．得られたクラスタが 3 個だった場合は 1 個のクラスタに 500 件以上の質問が含まれた場合とした．表 5 の中で，クラスタ C1 では「パソコン、周辺機器」，クラスタ C2 では「妊娠、出産」，クラスタ C3 では「話題の人物」，クラスタ C4 では「恋愛相談、人間関係の悩み」が最も質問数が多かった．本章ではこれら 4 つのカテゴリでの質問者と回答者の特徴を考察する．

「パソコン、周辺機器」ではクラスタ C1 の質問者の丁寧な記述に対し，回答者も丁寧な記述で返す文末表現のパターンが最も多かった．実際の質問では「このようなことを聞いて恐縮ですが」と少し調べれば分かりそうだが，調べ方が分からないので自分の質問の至らなさを申し訳なく思っていることを書いた上で，質問を丁寧に記述していた．これに対する回答では非難する回答者も居たが，その中で質問者の気持ちを汲み取って，丁寧な記述で回答する回答者もいた．質問者は自分の気持ちに応えてくれたことで好意を持ち，その回答者の書いた回答を選んだと考えられる．したがって，丁寧な記述で質問するこ

表 5: 1 つのクラスタに 400 件以上の質問が含まれたカテゴリ

クラスタ C1	クラスタ C2
パソコン、周辺機器	**妊娠、出産**
コミック	芸能人、タレント
野球	Yahoo!知恵袋
ファッション	結婚
レシピ、調理法	芸術、文学、歴史
不動産、引っ越し	政治、社会問題
テレビ、ラジオ	バイク
国内	海外
一般教養	家事
新車	動物、植物、ペット
Yahoo!オークション	自動車
クラスタ C3	クラスタ C4
話題の人物	**恋愛相談、人間関係の悩み**
子育ての悩み	小・中学校、高校
子育て、出産	料理、グルメ、レシピ
マナー	健康、病気、ダイエット
交通、地図	言葉、文学
ゲーム	
病気	
インターネット	

とは，回答者に対して申し訳なく思っている，という態度の表れであり，それに対し丁寧な記述で回答することは，質問者の気持ちを汲み取ったという態度の表れであると解釈できる．

「妊娠、出産」ではクラスタ C2 の質問者が意見や感想を記述することに対し，回答者は協調する意見や感想を返す文末表現のパターンが最も多かった．実際の質問では「自分はこう思うのだが，周りは自分と違うことを思っている」と，自分の意見と周りの意見のどちらを採用すれば良いのかを悩んでいる旨が質問されていた．これに対する回答では，質問者が正しいというもの，周りの人が正しいというものの両方があった．質問者は，妊娠，出産の不安の中で，自分を勇気づける言葉が欲しいと思っていると予想され，自分が正しいと言ってくれた回答者に好意を持ち，その回答者が書いた回答を選んだと考えられる．したがって，自分の意見や感想について質問することは，自分の考えの正当性を皆に評価して欲しいという態度の表れであり，それに対し，協調的な意見や感想を回答することは，不安を感じる質問者を勇気づけようとする回答者の態度の表れであると解釈できる．

「話題の人物」ではクラスタ C3 の質問者が答を問うことに対し，回答者は答のみを返す文末表現のパターンが最も多かった．実際の質問では,「○○の彼女は誰？」と特定の人物の詳細が質問されていた．これに対する回答では，その人物の特徴のみ記述されたもの，詳細な説明を加えて特徴が記述されたものなどがあった．質問者は簡潔な答だけを求めており，自分の要求に適した答のみを回答した回答者に好意を持ち，その回答者の回答を選んだと考えられる．したがって，簡潔に答を問う質問をするのは，余分な情報は不要であるという態度の表れであり，それに対し，答だけを回答することは質問者の求めることを汲み取ったという回答者の態度の表れであると解釈できる．

「恋愛相談、人間関係の悩み」ではクラスタ C4 の質問者が意見や感想を記述することに対し，回答者は異なる観点からの意見や感想を返す文末表現のパターンが最も多かった．実際の質問では「自分はこう思うのだが，どうすれば良いか分からない」と，直面している問題を解決する手法が見つからず悩んでいる旨が質問されていた．これに対する回答では，質問者が言うように問題をとらえ，解決する方法を提案している回答もあったが，その問題設定が間違っていると指摘する回答者もいた．質問者は回答者によって提示された解決方法も考えた上で，どうしたらよいかと質問したと予想される．その裏の意図を読み取った回答者は，全く別の観点から問題を考えてはどうかと，質問者の意見とは異なる観点からの意見を回答しており，質問者はその回答者の書いた回答を選んでいた．したがって，意見や感想を質問することは，クラスタ C2 の自分に同意して欲しいという態度の他に，裏の意図を読み取って欲しいという態度の表れもあり，異なる観点からの意見や感想を記述する回答者は質問者の裏の意図を読み取ったという態度の表れであると解釈できる．

5. 結論

知識共有コミュニティでの質問と回答の文末表現のパターンをカテゴリごとに調査した．本稿では文末表現のパターンを「質問とベストアンサーにのみ含まれる文末表現」と定義し，知識共有コミュニティの一つである Yahoo!知恵袋に投稿された質問と回答から，文末表現のパターンを計算機によって抽出し，人手による分析を行った．分析の結果，Yahoo!知恵袋では文末表現のパターンが 4 種類存在することが分かった．今後は，他の知識共有コミュニティの文末表現のパターンを解釈し，コミュニティの違いを明らかにしていきたい．

謝辞

ヤフー株式会社様には Yahoo!知恵袋の質問，回答の使用許可を頂戴しました．また，神戸学院大学の三浦麻子准教授から貴重な助言を頂戴しました．ここに記して感謝いたします．

参考文献

[Byrne71] D. Byrne：The attraction paradigm，Academic Press，1971.

[Core97] M. G. Core and J. F. Allen: Coding dialogs with the DAMSL annotation scheme, in Proc. of the American Association for Artificial Intelligence Fall Symposium

[Hearst93] M. A. Hearst and C. Plaunt, Subtopic structuring for full-length document accessm, in Proc. of the 16th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp.59–68,1993.

[Hovy00] E. Hovy, L. Gerber, U. Hermjakob, M. Junk, and C.-Y. Lin, Question Answering in Webclopedia, in Proc. of the 9th Text Retrieval Conf., pp.655–664, 2000.

[Jurafsky97] D. Jurafsky, E. Shriberg, and D. Biasca: Switchboard SWBD-DAMSL shallow-discourse-function annotation (coders manual, draft 13), Technical Report 97-02, University of Colorado, Institute of Cognitive Science, 1997.

[松本 97] 松本裕治，北内啓，山下達雄，平野善隆，今一修，今村友明：日本語形態素解析システム「茶筌」version1.0 使用説明書，NAIST Technical Report，NAIST-IS-TR97007，1997.

[徳久 06] 徳久良子，寺嶌立太：雑談における発話のやりとりと盛り上がりの関連，人工知能学会論文誌，Vol.21, No.2, pp.133–142, 2006.

[Vapnik95] V. N. Vapnik：The Nature of Statistical Learning Theory，Springer，1995.

[山下 02] 山下直美，石田亨，野村早恵子，早水哲雄：電子メールを用いた組織間交渉事例の分析，情報処理学会論文誌，Vol.43, No.1, pp.3355–3363, 2002.