質問回答サイトに投稿された質問回答文の因子得点の推定に N-gram を適用した分析手法

Analysis method of applying N-gram for estimating factor scores of statements posted at Q&A site

横山 友也 [1*]
Yuya Yokoyama[1*]

[1] 東京都立産業技術大学院大学 Advanced Institute of Industrial Technology
*Corresponding author: Yuya Yokoyama, yokoyama_yuya@aiit.ac.jp

Abstract For the purpose of eliminating mismatches between the questioners and respondents at Question and Answer (Q&A) sites, nine factors of impressions for Q&A statements have been experimentally obtained. Then through multiple regression analysis, factor scores have been estimated by using the feature values of statements e.g., syntactic information, etc. The factor scores estimated and obtained have been subsequently utilized for detecting respondents who are expected to appropriately answer a posted new question. This method, however, has greatly been dependent on the syntactic information extracted through morphological analysis. Moreover, this method has resulted in a number of explanatory variables and complicated multiple regression equation to estimate factor scores. On, the other hand, N-gram is known as an alternative syntactic analysis of morphological analysis. Therefore, in this paper, N-gram has been employed for estimating factor scores instead of syntactic information extracted through morphological analysis. In the analysis, N has been set from 2 to 5 (2-gram, 3-gram, 4-gram and 5-gram). Similar as the previous analysis, through multiple regression analysis, factor scores of nine factors are set as respondent variable. Meanwhile, fifty-eight feature values including N-gram are used as explanatory variable. The analysis result has conveyed that N-gram has shown as good estimation accuracy as morphological analysis. It has been also shown that the application of N-gram could influence the estimation of factor scores from the viewpoint of standardized partial regression coefficient.

Keywords Q&A site; N-gram; multiple regression analysis; factor score; standardized partial regression coefficient

1 はじめに

インターネットの急速な普及に伴って様々なウェブサービスが提供されており、その中でも質問回答サイトの利用者が年々増加の一途を辿っている[1]。質問回答サイトとは、インターネット上でユーザ同士が互いに質問と回答を投稿しあうコミュニティの一形態であり、種々の悩み事・相談事を解決する場であると同時に、膨大な知識が蓄積されたデータベースとして活用されるようになってきている。ソーシャル・ネットワーキング・サービス（SNS）とともに利用者が急増しており、その適切な運営は社会的に重要である。あるユーザが質問を投稿すると、他のユーザがその質問に対して回答を投稿する。質問文に対して最も適切と判断した回答文を質問者が「ベストアンサー」（BA）に選定し、その回答を行った回答者に謝礼として手持ちのポイントを贈与する。BA とは、質問文に対する満足度が最も高いと質問者が主観的に判断した回答文である。

しかし、質問文が投稿されても、その質問文が必ずしも適切な回答者の目に留まり回答されるわけではない。また、質問回答サイトは社会の知恵袋となってきているが、必ずしも正しい回答が蓄積されているわけではなく、誤答も多く発見される。適切な回答が得られないことで損害を被る状況が多発しており、社会的に大きな問題となってきている。従って、質問文に対して適切に回答できる利用者を求めることは、適切な回答を質問者に返すとともに正しい知識を蓄積するという点で非常に重要である。

筆者は、これまでの研究で、Yahoo! 知恵袋に投稿された 4 大ジャンル（オークション、パソコン、恋愛相談・人間関係、政治・社会問題）の質問回答文 12 組計 60 個の文章に対して印象評価実験を行った[2]。その結果に対して因子分析を施したところ、文章内容を表す 9 因子が得られた。また、全ての質問回答文の因子得点を推定可能にすることを目的として、文章の特徴量から文章の因子得点の推定を重回帰分析により試みた[3]。分析の結果、全般的に良好な推定精度が得られた。さらに、新規質問文に適切な回答を施すことが可能な回答者を探索する可能性を検証した。検証の結果、質問回答文間の距離と距離上位の出現回数は、質問文に適切に回答できるユーザの選択に役立てる可能性を示した[4]。この観察に基づき、出現回数と距離に基づくスコアと距離上位の出現回数に応じて、適切な回答者を決定する手法を提案した。提案手法を、スコアの平均値に基づく手法ならびに距離に基づく手法と、適合率と再現率で比較評価を行った[4, 5]。評価の結果、提案手法は、他手法よりも良好に回答者を推薦できることを示し、その精度は質問回答文のジャンルを考慮することでより向上することを示した[5]。

ここまで論じた手法は、形態素解析を講じて抽出した構文情報に重きを置いた手法である。一方で、構文解析のもう一種の手法である N-gram を講じた手法もあり、検討の余地がある。そこで、本稿では、N-gram を文章の特徴量として質問回答文の因子得点の推定と分析について詳述する[7-9]。ここでは、構文情報の代わりに N-gram を特徴量として使用し、重回帰分析を講じて因子得点の推定を行った。 2-gram（N=2）から 5-gram（N=5）の場合まで分析を行った結果、どの場合でも良好な推定結果が得られており、形態素解析のみを用いた場合と比較すると同等程度の推定精度が得られている。また、説明変数の因子得点への影響力を標準偏回帰係数の観点から評価したところ、N-gram は一定の影響力があることを示している。

本稿の構成は次の通りである。第 2 節では、これまでの研究として、形態素解析に基づいた構文情報を用いた場合の因子得点の推定について説明する。第 3 節では、N-gram を特徴量として用いた場合の因子得点の推定について詳述する。最後に、第 4 節で本稿をまとめる。

2 形態素解析に基づく構文情報を使う場合の因子得点の推定

2.1 文章の因子の獲得

2005 年 9 月に Yahoo!知恵袋[1]に投稿された 12 組 60 個の 4

Received: 2023-10-29; Published online: 2024-02-29

大カテゴリー（Yahoo!オークション、パソコン、恋愛相談・人間関係、政治・社会問題）の質問回答文に対して、印象評価実験を実施した。実験結果に対して因子分析を施したところ、文章に関する因子が 9 個得られた[2]。因子とは、複数の印象語により説明された文章の性質を意味する。的確性、不快性、独創性、容易性、執拗性、曖昧性、感動性、努力性、熱烈性の 9 因子が得られた[2]。なお、因子の名称は、各因子に対応する印象語を包括的に表していると著者達が判断した名称を付与している。因子に対応する印象語を表 1 に示す。また、文章の特徴を表現するのに使用する因子得点も得られている[2]。

表 1　　9 因子と対応する印象語[2]

因子	印象語				
第1因子（的確性）	説得力がある	流暢な	重要な	美しい	好ましい
	真実味がある	巧みな	清々しい	妥当な	充実した
	素晴らしい	的確な	丁寧な		
第2因子（不快性）	非常識な	憤慨した	不快な	残念な	不当な
	幻滅した	呆れる	怖い		
第3因子（独創性）	独創的な	予想外な	特殊な	斬新な	不思議な
第4因子（容易性）	易しい	明瞭な	難しい		
第5因子（執拗性）	細かい	しつこい	長い		
第6因子（曖昧性）	曖昧な	不充分な			
第7因子（感動性）	心温まる	感動的な			
第8因子（努力性）	涙ぐましい				
第9因子（熱烈性）	熱い	力強い			

2.2　形態素解析に基づく文章の特徴量を用いた因子得点の推定

2.1 節で得られた因子得点は、実験で使用した質問回答文 60 件から得られたものだけである。そこで、任意の文章の因子得点の推定も可能とするために、文章の特徴量に対し重回帰分析を講じた。ここで、構文解析の一種である形態素解析を通して抽出した文章の特徴量を用いて分析を行った[3]。分析に使用した文章の特徴量 77 個（g1～g77）を表 2 に示す。以下、各特徴量を簡潔に説明する。

- 構文情報（g1 - g36）：文章の数や長さ、名詞や動詞等の品詞の数や割合、といった構文情報を抽出した。感嘆符や疑問符等の具体的な記号も特徴量として使用している[3]。なお、g18 の TTR は Type Token Ratio の略記で、「ある文章における語彙の豊富さを示す指標」であり、文章の総語数に対する語彙数の比率を表している[3]。
- 単語心像性（g37 - g38）：単語から喚起されるイメージが、どの程度思い浮かべやすいかを示す主観的特性である[3]。
- 文末表現（g39 - g64）：「ぞ」「だ」「よ」「ね」「か」「な」「し」「です」「ます」「たい」「ない」を使用している[3]。
- 単語親密度（g65 - g71）：単語にどの程度なじみがあるかを表す指標である[3]。
- 表記妥当性（g72 - g77）：単語表記の妥当性を表す指標である[3]。

印象評価実験で使用された 60 個の質問回答文に関して、表 2 に示した計 77 個の説明変数を基盤とした 281 個の二次項（説明変数同士の積）を説明変数とし、表 2 で示した 9 因子の因子得点を目的変数として、重回帰分析を施した[3]。重回帰式が結果として得られた。例えば、第 5 因子（執拗性）の因子得点 $\boldsymbol{y_5}$ の推定式は式(1)により表される[3]。

$$\begin{aligned} y_5 = {} & 0.182 g_1 g_{18} + 0.000280 g_5 g_6 + 0.00467 g_{24} g_{60} \\ & - 0.0467 g_{23} g_{58} + 0.00985 g_4 g_{58} + 0.102 g_{23} g_{29} \\ & + 0.339 g_{33} g_{45} - 0.201 g_{66} - 0.0149 g_{51} g_{72} \\ & - 0.266 g_{51} g_{54} - 0.672 \end{aligned} \qquad (1)$$

表 2　　文章の特徴量[3]

(a)　構文情報

g	特徴量	g	特徴量
g1	助動詞（語彙数）	g19	全角記号(%)
g2	接頭詞	g20	英数字(%)
g3	記号（語彙数）	g21	全角英数字(%)
g4	文数	g22	名詞(%)
g5	文の長さ平均（字数）	g23	形容詞(%)
g6	カタカナ（語数）	g24	副詞(%)
g7	全角記号（語数）	g25	連体詞(%)
g8	全角英数字（語数）	g26	接続詞(%)
g9	形容詞（語数）	g27	感動詞(%)
g10	副詞（語数）	g28	「!」の数
g11	連体詞（語数）	g29	「?」の数
g12	接続詞（語数）	g30	句点の数
g13	感動詞（語数）	g31	読点の数
g14	ひらがな（%）	g32	中点の数
g15	漢字（%）	g33	3点リーダの数
g16	カタカナ（%）	g34	鍵括弧の数
g17	記号（%）	g35	括弧の数
g18	TTR	g36	「/」の数

(b)　単語心像性

g	特徴量	g	特徴量
g37	単語心像性4点台（語数）	g38	単語心像性6.5以上7.0未満（語数）

(c)　文末表現

g	特徴量	g	特徴量
g39	か（語数）	g52	ぞ（%）
g40	な（語数）	g53	だ（%）
g41	し（語数）	g54	よ（%）
g42	たい（語数）	g55	ね（%）
g43	ない（語数）	g56	か（%）
g44	だ（文末語数）	g57	です（%）
g45	か（文末語数）	g58	ます（%）
g46	な（文末語数）	g59	ない（%）
g47	し（文末語数）	g60	か（文末%）
g48	です（文末語数）	g61	ですか（語数）
g49	ます（文末語数）	g62	ないです（語数）
g50	たい（文末語数）	g63	ますか（語数）
g51	ない（文末語数）	g64	ました（語数）

(d)　単語親密度

g	特徴量	g	特徴量
g65	単語親密度該当単語率	g69	単語親密度5.5以上6.0未満（語数）
g66	単語親密度6.5以上7.0未満（語彙数）	g70	単語親密度6点台（語数）
g67	単語親密度4点台（語数）	g71	単語親密度6.0以上6.5未満（語数）
g68	単語親密度5点台（語数）		

(e)　表記妥当性

g	特徴量	g	特徴量
g72	表記妥当性該当単語率	g75	表記妥当性4点台（語数）
g73	表記妥当性3点台（語数）	g76	表記妥当性4.0以上4.5未満（語数）
g74	表記妥当性3.5以上4.0未満（語数）	g77	表記妥当性5点台（語数）

$\boldsymbol{g_k(1 \le k \le 77)}$ は説明変数を意味している。紙面が限られているため、全ての重回帰式の組をここには掲載しない。ここで、第 1 因子には的確性という名前を付けており、本論文に特に関連すると考えられるが、第 1 因子の因子得点は、主に、接続詞の語数、「ます」の割合、読点の数、高い単語親密度をもつ単語の語数によって推定され[2]、丁寧さや好ましさ等も含んだ因子と考えられる。これは、表 1 に示した第 1 因子を説明する印象語と合致する。的確性という名前を付けているが、これら

の印象も含むことに注意されたい。推定精度の良好性を示す重相関係数を表 3 に示す。9 因子全ての値が 0.9 以上であるため、9 因子とも推定精度が非常に良好であるといえる。

表 3　重相関係数（構文情報）[3]

因子	重相関係数
第1因子（的確性）	0.989
第2因子（不快性）	1.000
第3因子（独創性）	0.999
第4因子（容易性）	1.000
第5因子（執拗性）	0.925
第6因子（曖昧性）	1.000
第7因子（感動性）	0.963
第8因子（努力性）	0.950
第9因子（感動性）	1.000

2.3　因子得点に基づく適切な回答者の選出手法と評価

2.2 節で求めた重回帰分析をもとに、実験に使用していない任意の質問回答文について因子得点を算出して、新規質問文に適切に回答できることが予期される回答者を探索する可能性を検証した[4]。検証の結果、質問回答文間の距離と距離上位の出現回数は、質問文に適切に回答できるユーザの選択に役立てる可能性を示した[4]。この検証結果に基づいて、出現回数と距離に基づくスコアと距離上位の出現回数に応じて、適切な回答者を選出する手法を提案した[4, 5]。質問回答文 3 組（1 組の構成：質問文 1 件・質問文に非対応の回答文 100 件）を用いた評価実験を通じて、提案手法をスコアの平均値に基づく手法ならびに距離に基づく手法と、適合率と再現率で比較評価を行った。評価の結果、提案手法は他手法よりも良好に回答者を推薦可能であることを示し[4]、その推定精度は質問回答文のジャンルを考慮することで向上することを示した[5]。

3　N-gram を特徴量として用いた場合の因子得点の推定

3.1　目的

2 節で述べた分析手法は、主に形態素解析を通じて抽出された文章の特徴量を使用した手法である。しかし、表 2 に示したように、使用した文章の特徴量 77 種のうち 36 種が形態素解析に基づいた構文情報であることから、この手法は形態素解析に重きを置いた手法となっている。また、因子得点を推定するための重回帰式も複雑なモデルとなっている。一方で、形態素解析と同様に構文解析として知られている方法として N-gram がある。そこで、本稿では、形態素解析を主に特徴量に使用した場合と同様の分析手法で、N-gram を特徴量して用いた場合に因子得点の推定が可能であるかどうかを検証する[6-8]。

3.2　N-gram

N-gram とは、テキスト内における N 個単位の文字や形態素、ないしは品詞の連鎖のことである[9, 10]。N は 2 以上の任意の整数が入るが、一般的には N=2 または N=3 が適用され、それぞれバイグラム（bigram）、トリグラム（trigram）と称される。たとえば、テキスト内における「今日は」を例にすると、文字 3-gram であれば「今　日　は」のような 3 文字の連鎖、単語 2-gram であれば「今日　は」のような 2 単語の連鎖、品詞 2-gram であれば「名詞　助詞」のような 2 品詞の連鎖をそれぞれ表している。このように、N-gram は隣接する要素を機械的に抽出する手法であり、様々な分野に適用されている[9]。文字 N-gram は、形態素解析を要することなく集計することが可能であり、形態素解析における分かち書きの誤り等の影響を受けずにテキストを分析することが可能である[10]。さらに、品詞 N-gram は、文章を品詞の単位に抽象化するため、文章の内容の影響に左右されることなく文章の構造を捉えることができる利点がある[10]。したがって、本研究では、60 件の質問回答文の品詞 N-gram を文章の特徴量として使用する。

2.1 節で使用した 60 件の質問回答文のうち、1 つの質問文に品詞 2-gram を適用した例を表 4 に示す。便宜として、表 4-(a) に例として示した質問文を QA04 と表す。表 4-(b)は、QA04 の品詞 2-gram の例と出現数について、出現数の降順で示している。出現数が 2 以上の品詞 2-gram に関しては、それぞれの一例のみを「例」の項に示している。

表 4　品詞 2-gram の例 [6-8]

(a)文章の例（QA04）

QA04	パソコン初心者です。デジカメで撮った画像を プリントアウトしたところ画像が暗いのですが、 明るくする方法をご存知の方回答をお願いします。

(b)2-gram の適用例ならびに出現数

2-gram	例	出現数
[名詞 - 助詞]	[画像 - を]	6
[名詞 - 名詞]	[パソコン - 初心者]	4
[助詞 - 名詞]	[の - 方]	4
[動詞 - 助動詞]	[する - ます]	3
[助動詞 - 記号]	[ます - 。]	2
[助動詞 - 名詞]	[た - ところ]	2
[名詞 - 助動詞]	[初心者 - です]	2
[名詞 - 動詞]	[お願い - する]	2
[記号 - 形容詞]	[、 - 明るい]	1
[記号 - 名詞]	[。 - デジカメ]	1
[形容詞 - 動詞]	[明るい - する]	1
[形容詞 - 名詞]	[暗い - の]	1
[助詞 - 記号]	[が - 、]	1
[助詞 - 形容詞]	[が - 暗い]	1
[助詞 - 動詞]	[で - 撮る]	1
[助動詞 - 助詞]	[です - が]	1
[動詞 - 名詞]	[する - 方法]	1

3.3　N-gram の特徴量

本研究では、統計解析フリーソフト R[11]において、RMeCab のパッケージを用いて 品詞 N-gram の特徴量を抽出した。本稿では、2-gram（N=2）から 5-gram（N=5）までの場合について分析を行った。品詞 3-gram の場合を例にして、表 5 に示す特徴量の選択方法を説明する。まず、60 件の質問回答文から、

表 5　質問回答文 60 件の N-gram の選択方法の概要

(a) ソート前

	Q&A	…	[名詞-名詞-名詞]	[名詞-名詞-助動詞]	…	[名詞-動詞-名詞]	[記号-名詞-助詞]	…
			85組					
Q&A60件	QA01	…	0	0	…	0	3	…
	AA01-01	…	0	0	…	0	8	…
	AA01-02	…	1	0	…	0	5	…
	AA01-03	…	0	0	…	0	3	…
	AA01-04	…	0	0	…	0	1	…
	QA02	…	1	0	…	2	1	…
	…	…	…	…	…	…	…	…
	QA12-03	…	0	0	…	0	0	…
	QA12-04	…	6	2	…	0	8	…
	3-gram計	…	153	20	…	20	162	…

(b) ソート後

Q&A	2gr_g1 [記号-名詞-助詞]	2gr_g2 [名詞-名詞-名詞]	…	2gr_g16 [名詞-動詞-名詞]	2gr_g17 [名詞-名詞-助動詞]
QA01	3	0	…	0	0
AA01-01	8	0	…	0	0
AA01-02	5	1	…	0	0
AA01-03	3	0	…	0	0
AA01-04	1	0	…	0	0
QA02	1	1	…	2	0
…	…	…	…	…	…
QA12-03	0	0	…	0	0
QA12-04	8	6	…	0	2
3-gram計	162	153	…	20	20

のべ 85 種類の品詞 3-gram が生成される（表 5-(a)：ソート前の状態）。それぞれの品詞 3-gram の合計出現数を求めた上で、合計出現数の降順に品詞 3-gram を並べる（表 5-(b); ソート後の状態）。分析に用いる 3-gram の特徴量数は、試験的に 17 個と定めて分析を行った。同様に、2-gram, 4-gram, 5-gram の場合もそれぞれ特徴量を 17 個として分析を行った。N-gram の特徴量を表 6～表 9 にそれぞれ示す。表記について、N-gram の特徴量を Ngr_g1, Ngr_g2, …, Ngr_g17 と表している。したがって、2-gram を例とすると、2gr_g1, 2gr_g2, …, 2gr_g17 と表す。3-gram, 4-gram, 5-gram についても、2-gram と同様の表記に準じる。

表 6　2-gram の特徴量［6］

2-gram	特徴量	2-gram	特徴量
2gr_g1	[名詞-助詞]	2gr_g10	[名詞-助動詞]
2gr_g2	[助詞-動詞]	2gr_g11	[助動詞-記号]
2gr_g3	[助詞-名詞]	2gr_g12	[動詞-名詞]
2gr_g4	[名詞-名詞]	2gr_g13	[名詞-動詞]
2gr_g5	[記号-名詞]	2gr_g14	[助動詞-名詞]
2gr_g6	[動詞-助動詞]	2gr_g15	[助動詞-助動詞]
2gr_g7	[助詞-記号]	2gr_g16	[記号-記号]
2gr_g8	[動詞-助詞]	2gr_g17	[助詞-助詞]
2gr_g9	[助動詞-助詞]		

表 7　3-gram の特徴量［6］

3-gram	特徴量	3-gram	特徴量
3gr_g1	[記号-名詞-助詞]	3gr_g10	[助動詞-助動詞-記号]
3gr_g2	[名詞-名詞-名詞]	3gr_g11	[記号-記号-記号]
3gr_g3	[助詞-記号-名詞]	3gr_g12	[助詞-記号-副詞]
3gr_g4	[助詞-動詞-名詞]	3gr_g13	[名詞-動詞-助動詞]
3gr_g5	[記号-名詞-名詞]	3gr_g14	[助動詞-名詞-助動詞]
3gr_g6	[動詞-名詞-助詞]	3gr_g15	[記号-名詞-記号]
3gr_g7	[名詞-助動詞-助詞]	3gr_g16	[名詞-動詞-名詞]
3gr_g8	[動詞-助動詞-名詞]	3gr_g17	[名詞-名詞-助動詞]
3gr_g9	[名詞-助詞-形容詞]		

表 8　4-gram の特徴量［7］

4-gram	特徴量	4-gram	特徴量
4gr_g1	[名詞-助詞-名詞-助詞]	4gr_g10	[助動詞-記号-名詞-助詞]
4gr_g2	[名詞-名詞-名詞-名詞]	4gr_g11	[助詞-記号-名詞-助詞]
4gr_g3	[名詞-助詞-動詞-助詞]	4gr_g12	[名詞-助詞-動詞-名詞]
4gr_g4	[記号-名詞-助詞-名詞]	4gr_g13	[名詞-名詞-助詞-名詞]
4gr_g5	[助詞-名詞-助詞-動詞]	4gr_g14	[助詞-動詞-名詞-助詞]
4gr_g6	[名詞-助詞-動詞-助動詞]	4gr_g15	[名詞-助詞-名詞-名詞]
4gr_g7	[助詞-動詞-助詞-動詞]	4gr_g16	[名詞-助詞-名詞-動詞]
4gr_g8	[助詞-名詞-助詞-名詞]	4gr_g17	[動詞-助詞-動詞-助動詞]
4gr_g9	[記号-名詞-助詞-動詞]		

表 9　5-gram の特徴量［8］

5-gram	特徴量	5-gram	特徴量
5gr_g1	[名詞-助詞-名詞-助詞-動詞]	5gr_g10	[助動詞-助詞-記号-名詞-助詞]
5gr_g2	[名詞-助詞-動詞-助詞-動詞]	5gr_g11	[名詞-名詞-助詞-名詞-助詞]
5gr_g3	[名詞-助詞-名詞-助詞-名詞]	5gr_g12	[助詞-記号-名詞-助詞-動詞]
5gr_g4	[記号-名詞-助詞-名詞-助詞]	5gr_g13	[動詞-助動詞-記号-名詞-助詞]
5gr_g5	[助詞-名詞-助詞-動詞-助詞]	5gr_g14	[名詞-助詞-動詞-助動詞-助詞]
5gr_g6	[助詞-名詞-助詞-名詞-助詞]	5gr_g15	[名詞-助詞-名詞-名詞-助詞]
5gr_g7	[助動詞-記号-名詞-助詞-名詞]	5gr_g16	[助詞-記号-名詞-名詞-助詞]
5gr_g8	[名詞-助詞-動詞-名詞-助詞]	5gr_g17	[助詞-記号-名詞-助詞-名詞]
5gr_g9	[助詞-動詞-助詞-動詞-助動詞]		

3.4　分析手法

2.2 節で示した手法と同様に、N-gram を用いた文章の特徴量を用いて重回帰分析を施して因子得点の推定を試みる。2.2 節の場合と同様に、質問回答文 60 件の因子得点を目的変数に設定する。一方で、2.2 節での分析と異なる点として、形態素解析に基づいた構文情報（36 種：g1 - g36）に代わって、N-gram に基づいた特徴量（17 種：Ngr_g1 - Ngr_g17）を新規特徴量として説明変数に用いる[6-8]。一方で、2.2 節で使用した単語心像性（g37 – g38）・文末表現（g39– g64）・単語親密度（g65 – g71）・表記妥当性（g72 – g77）の計 41 種については引き続き説明変数として使用する。したがって、計 58 種の特徴量を説明変数として設定する。使用する説明変数の一覧ならびに形態素解析（構文情報）の場合と N-gram の場合との差異を図 1 に示す。

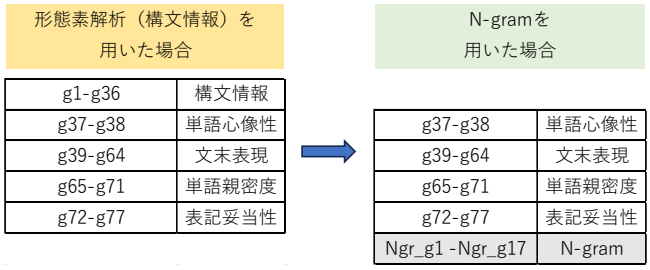

図 1　　使用する文章の特徴量（左：形態素解析、右：N-gram）

3.5　分析結果

3.5.1　重相関係数

重回帰分析の結果として、2.2 節と同様に、推定精度の指標を表す重相関係数が得られた。N-gram の重相関係数を表 10 に示す。表 10 の結果から、どの N-gram においても重相関係数が 0.9 以上になっており、全般に良好な推定精度が得られている[6-8]。表 3 の形態素解析を用いた場合の結果と比較すると、形態素解析を用いた場合と同等程度の推定精度が得られている。

表 10　重相関係数（N-gram）[8]

因子	2-gram	3-gram	4-gram	5-gram
第1因子　的確性	0.989	0.993	0.999	0.998
第2因子　不快性	0.999	0.987	0.985	0.991
第3因子　独創性	0.981	0.998	0.971	0.976
第4因子　容易性	0.990	0.995	0.993	0.994
第5因子　執拗性	0.993	0.976	0.994	0.999
第6因子　曖昧性	0.998	0.994	0.983	0.986
第7因子　感動性	0.999	0.996	0.945	0.992
第8因子　努力性	0.995	0.968	0.988	1.00
第9因子　熱烈性	0.995	0.998	0.973	0.954

3.5.2　標準偏回帰係数

説明変数（文章の特徴量）の目的変数（因子得点）への影響力の大きさを調べるため、各因子について標準偏回帰係数（Standardized Partial Regression Coefficient = SPRC）の絶対値 1.0 以上の変数を調べた。該当する変数のうち、上位 3 件で SPRC が大きな変数を正と負それぞれで上位最大 3 件を調べた。2-gram, 3-gram, 4-gram, 5-gram についてそれぞれの結果を表 11～表 14 に示す。但し、3-gram の第 5 因子（表 12）と 4-gram の第 6 因子（表 13）に関して、SPRC の絶対値が 1.0 以上の変数が存在しないため、絶対値 1.0 未満の範囲で正と負それぞれで最も絶対値が大きな変数を示す。「特徴量」の項は、「説明変数」の項に対応する特徴量の種類を表している。特徴量の種類とは、N-gram (Ngr_g1 -Ngr_g17)、単語心像性 (Word Imageability = WI; g37-g38)、文末表現 (Closing sentence expression=Closing; g39-g64)、単語親密度 (Word Familiarity = WF; g65-g71)、表記妥当性 (Notation Validity = NV; g72-g77)のいずれかである。

表 11　SPRC の絶対値が 1.0 以上の変数（2-gram）[6]

第1因子（的確性）

説明変数	特徴量	SPRC
2gr_g7	2-gram	1.27
2gr_g10	2-gram	1.24

第2因子（不快性）

説明変数	特徴量	SPRC
2gr_g2	2-gram	5.65
g70	WF	3.28
2gr_g9	2-gram	2.05
g39	Closing	-2.60
2gr_g10	2-gram	-2.86
2gr_g6	2-gram	-3.09

第3因子（独創性）

説明変数	特徴量	SPRC
g65	WF	3.68
g39	Closing	3.01
2gr_g8	2-gram	2.83
g72	NV	-2.18
g70	WF	-3.49
2gr_g2	2-gram	-6.90

第4因子（容易性）

説明変数	特徴量	SPRC
2gr_g3	2-gram	3.71
2gr_g2	2-gram	2.82
2gr_g13	2-gram	-1.17
2gr_g1	2-gram	-5.30

第5因子（執拗性）

説明変数	特徴量	SPRC
2gr_g3	2-gram	3.00
2gr_g6	2-gram	1.59
g45	Closing	1.35
2gr_g9	2-gram	-1.23
2gr_g1	2-gram	-2.40

第6因子（執拗性）

説明変数	特徴量	SPRC
2gr_g1	2-gram	1.66
2gr_g9	2-gram	1.29
2gr_g2	2-gram	-1.89
2gr_g3	2-gram	-2.32

第7因子（感動性）

説明変数	特徴量	SPRC
2gr_g1	2-gram	4.12
2gr_g10	2-gram	2.93
2gr_g8	2-gram	2.19
2gr_g3	2-gram	-2.43
2gr_g9	2-gram	-3.17
2gr_g2	2-gram	-5.20

第8因子（努力性）

説明変数	特徴量	SPRC
2gr_g6	2-gram	2.85
g45	Closing	2.25
2gr_g1	2-gram	2.20
g65	WF	-1.41
2gr_g2	2-gram	-1.61
g37	WI	-1.71

第9因子（熱烈性）

説明変数	特徴量	SPRC
2gr_g10	2-gram	5.21
2gr_g6	2-gram	4.54
2gr_g8	2-gram	3.54
2gr_g13	2-gram	-2.98
g70	WF	-5.90
2gr_g2	2-gram	-12.08

表 12　SPRC の絶対値が 1.0 以上の変数（3-gram）[6]

第1因子（的確性）

説明変数	特徴量	SPRC
g70	WF	1.90
g37	WI	1.47
g43	Closing	1.14
g62	Closing	-1.06
3gr_g6	3-gram	-1.54
g64	Closing	-1.65

第2因子（不快性）

説明変数	特徴量	SPRC
g76	NV	1.16
g73	NV	-0.77

第3因子（独創性）

説明変数	特徴量	SPRC
g64	Closing	2.83
3gr_g6	3-gram	2.03
g65	WF	1.95
g76	NV	-1.68
g37	WI	-2.09
g70	WF	-2.36

第4因子（容易性）

説明変数	特徴量	SPRC
g65	WF	1.42
g44	Closing	1.29
g70	WF	-1.10
g76	NV	-1.34
g72	NV	-1.72

第5因子（執拗性）

説明変数	特徴量	SPRC
g45	Closing	0.97
g60	Closing	-0.47

第6因子（執拗性）

説明変数	特徴量	SPRC
g66	WF	1.04
3gr_g4	3-gram	-1.22
g43	Closing	-1.31
g70	WF	-1.34

第7因子（感動性）

説明変数	特徴量	SPRC
3gr_g6	3-gram	1.56
3gr_g4	3-gram	-1.30

第8因子（努力性）

説明変数	特徴量	SPRC
3gr_g6	3-gram	2.08
g68	WF	1.61
g59	Closing	1.59
3gr_g13	3-gram	-1.37
g43	Closing	-1.58
g37	WI	-1.61

第9因子（熱烈性）

説明変数	特徴量	SPRC
g66	WF	2.04
g65	WF	1.87
g73	NV	1.77
g76	NV	-2.04
3gr_g4	3-gram	-2.11
g70	WF	-3.26

表 13　SPRC の絶対値が 1.0 以上の変数（4-gram）[7]

第1因子（的確性）

説明変数	特徴量	SPRC
g37	WI	3.76
g68	WF	3.15
4gr_g8	4-gram	2.15
4gr_g7	4-gram	-1.91
4gr_g16	4-gram	-2.01
g76	NV	-3.12

第2因子（不快性）

説明変数	特徴量	SPRC
4gr_g6	4-gram	1.45
g56	Closing	1.33
4gr_g7	4-gram	1.32
g39	Closing	-1.04

第3因子（独創性）

説明変数	特徴量	SPRC
g76	NV	4.07
4gr_g9	4-gram	2.58
4gr_g16	4-gram	2.35
4gr_g8	4-gram	-2.88
g68	WF	-4.00
g37	WI	-4.51

第4因子（容易性）

説明変数	特徴量	SPRC
4gr_g9	4-gram	1.56
g76	NV	1.22
4gr_g5	4-gram	1.19
g62	Closing	-1.04
4gr_g1	4-gram	-1.08
4gr_g6	4-gram	-1.39

第5因子（執拗性）

説明変数	特徴量	SPRC
4gr_g8	4-gram	1.73
4gr_g5	4-gram	1.50
g43	Closing	-1.17
4gr_g1	4-gram	-1.80

第6因子（執拗性）

説明変数	特徴量	SPRC
g62	Closing	0.52
4gr_g6	4-gram	0.50
4gr_g14	4-gram	0.49
g44	Closing	-0.62
4gr_g9	4-gram	-0.64
g73	NV	-0.66

第7因子（感動性）

説明変数	特徴量	SPRC
g76	NV	4.33
g48	Closing	2.32
g72	NV	1.92
g37	WI	-2.36
g44	Closing	-2.40
g73	NV	-2.43

第8因子（努力性）

説明変数	特徴量	SPRC
g76	NV	4.07
g48	Closing	2.43
4gr_g16	4-gram	2.28
4gr_g8	4-gram	-2.76
g68	WF	-2.80
g37	WI	-4.55

第9因子（熱烈性）

説明変数	特徴量	SPRC
g65	WF	2.25
4gr_g1	4-gram	2.02
g44	Closing	1.25
4gr_g9	4-gram	-1.39
4gr_g4	4-gram	-1.44
g76	NV	-2.05

表 14　SPRC の絶対値が 1.0 以上の変数（5-gram）[8]

第1因子（的確性）		
説明変数	特徴量	SPRC
g37	WI	1.08
5gr_g6	5-gram	-1.10

第2因子（不快性）		
説明変数	特徴量	SPRC
g56	Closing	1.74
g49	Closing	1.04
g60	Closing	-1.10
g39	Closing	-1.86

第3因子（独創性）		
説明変数	特徴量	SPRC
g65	WF	3.89
g73	NV	2.12
5gr_g17	5-gram	1.67
g59	Closing	-1.62
g76	NV	-2.15
g72	NV	-2.56

第4因子（容易性）		
説明変数	特徴量	SPRC
g48	Closing	2.69
g76	NV	2.61
5gr_g5	5-gram	2.05
g68	WF	-1.99
5gr_g9	5-gram	-2.24
g37	WI	-3.62

第5因子（執拗性）		
説明変数	特徴量	SPRC
g75	NV	2.69
g76	NV	1.40
g61	Closing	1.29
g70	WF	-1.80
g73	NV	-2.12
g68	WF	-2.20

第6因子（執拗性）		
説明変数	特徴量	SPRC
g37	WI	1.76
g73	NV	1.25
g68	WF	1.22
g76	NV	-1.09
g75	NV	-1.44
g48	Closing	-1.61

第7因子（感動性）		
説明変数	特徴量	SPRC
g76	NV	2.63
g48	Closing	2.13
5gr_g12	5-gram	1.74
g37	WI	-2.06
g73	NV	-2.32
g44	Closing	-2.36

第8因子（努力性）		
説明変数	特徴量	SPRC
g72	NV	2.09
g48	Closing	1.81
5gr_g6	5-gram	1.47
g74	NV	-1.39
g37	WI	-2.05
g65	WF	-2.49

第9因子（熱烈性）		
説明変数	特徴量	SPRC
g68	WF	3.02
g73	NV	2.94
g70	WF	2.39
g61	Closing	-1.64
g75	NV	-2.38
g76	NV	-2.48

表 11 から表 14 の結果より、正または負に SPRC の絶対値の上位 3 件以内に N-gram の変数が該当する変数は、2-gram は 9 因子全て、3-gram は 6 因子、4-gram は 8 因子、5-gram は 5 因子である。これらの結果より、N-gram の因子得点への影響力が大きいことがわかる[6-8]。これらの結果より、因子得点の推定にあたり N-gram を文章の特徴量として考慮することには一定の有効性があるといえる。

3.6　考察

構文情報と N-gram とで重相関係数を比較しやすくするため、表 3（構文情報）と表 10（N-gram）の結果をまとめたものを表 15 に示す。重相関係数の観点からいえば、どの特徴量の場合でも因子得点の推定には実用的であるといえる結果が得られている。しかし、最良な因子得点の推定を求めるとするならば、重相関係数が最良となった結果の場合を適用すれば良いともいえる。表 15 の結果より、構文情報が最良の結果となったのは第 2 因子（不快性）、第 3 因子（独創性）、第 4 因子（容易性）、第 6 因子（曖昧性）、第 9 因子（熱烈性）の 5 因子である。一方、2-gram が最良の結果となったのは第 7 因子（感動性）、4-gram が最も良い数字となったのは第 1 因子（的確性）である。さらに、5-gram が最良の結果となったのは第 5 因子（執拗性）と第 8 因子（努力性）の 2 因子である。

表 15　重相関係数（形態素解析・N-gram）[3, 8]

因子	構文情報	N-gram			
		2-gram	3-gram	4-gram	5-gram
第1因子　的確性	0.989	0.989	0.993	0.999	0.998
第2因子　不快性	1.00	0.999	0.987	0.985	0.991
第3因子　独創性	0.999	0.981	0.998	0.971	0.976
第4因子　容易性	1.00	0.990	0.995	0.993	0.994
第5因子　執拗性	0.925	0.993	0.976	0.994	0.999
第6因子　曖昧性	1.00	0.998	0.994	0.983	0.986
第7因子　感動性	0.963	0.999	0.996	0.945	0.992
第8因子　努力性	0.950	0.995	0.968	0.988	1.00
第9因子　熱烈性	1.00	0.995	0.998	0.973	0.954

しかし、本研究ではこれまで文章の意味や内容を考慮していない。したがって、これらの要素を目的として、今後は意味解析を本手法に適用する必要性がある。

4　まとめ

本稿では、これまでの研究で主に使用してきた構文情報に代わって、N-gram を文章の特徴量として、重回帰分析を実施して質問回答文の因子得点の推定を行った。ここでは、構文情報の代わりに N-gram を特徴量として使用し、重回帰分析を講じて因子得点の推定を行った。2-gram から 5-gram まで分析を行った結果、すべての場合において良好な推定結果が得られており、構文情報を用いた場合と比べると同程度の推定精度が得られている。さらに、説明変数（文章の特徴量）が目的変数（因子得点）にどの程度影響を及ぼすかを調べるために、標準偏回帰係数の絶対値が 1.0 以上の変数に着目したところ、2-gram から 5-gram のどの場合においても、N-gram が絶対値の上位 3 件以内に該当する因子がのべ 36 因子中 28 因子（2-gram：9 因子、3-gram：6 因子、4-gram：8 因子、5-gram：5 因子）あることから、N-gram の考慮は因子得点に一定の影響力を及ぼしていることを示している。

今後の課題としては、文章の内容や意味を考慮する必要性がある。また、過学習の可能性を回避するためにクロス・バリデーションを適用して分析する必要性がある。さらに、2.3 節では、因子得点を用いて適切な回答者の選出手法について記述している[4-6]。ここで示した手法は形態素解析を通じて得られた構文情報に基づく手法であるので、N-gram に基づく手法で適切な回答者の選出手法が実施できるかどうかを検証する必要性がある。まずは、2-gram を特徴量とした場合から検証を進めていく方針である。また、本手法を他言語または他分野のデータセットに汎用できるかどうかを検証することも今後の課題の一つである。

謝辞

本研究は、一部、科研費（26008587, 20K19933）の助成を受けて行われたものである。また、国立情報学研究所から提供を受けた、Yahoo!知恵袋のデータを利用している。

参考文献

1. Yahoo! 知恵袋. [cited 7 Feb 2024]. Available: http://chiebukuro.yahoo.co.jp/.
2. 横山友也, 宝珍輝尚, 野宮浩揮, 佐藤哲司. 質問回答サイトの質問文と回答文の印象評価とベストアンサーの推定. 日本感性工学会論文誌. 2011;10;2: 221–230. Available: https://www.jstage.jst.go.jp/article/jjske/10/2/10_2_221/_pdf/-char/ja
3. Yokoyama Y., Hochin T., Nomiya H. Using Feature Values of Statements to Improve the Estimation Accuracy of Factor Scores of

Impressions of Question and Answer Statements. International Journal of Affective Engineering. 2013;13;1: 19–26. Available: https://www.jstage.jst.go.jp/article/ijae/13/1/13_19/_pdf/-char/ja
4. 横山友也, 宝珍輝尚, 野宮浩揮. 質問回答サイトにおける質問文への適切な回答者の選出法. 日本感性工学会論文誌. 2016;15;1: 21–29. Available: https://www.jstage.jst.go.jp/article/jjske/15/1/15_TJSKE-D-15-00033/_pdf/-char/ja/
5. Yokoyama Y., Hochin T., Nomiya H. Towards Detecting Appropriate Respondents to Questions Posted at Q&A Sites by Disregarding and Considering Categories of Answer Statements. International Journal of Affective Engineering. 2016;15;2: 167–175. Available: https://www.jstage.jst.go.jp/article/ijae/advpub/0/advpub_IJAE-D-15-00031/_pdf/-char/ja
6. Yokoyama Y., Hochin T., Nomiya H. Application of 2-gram and 3-gram to Obtain Factor Scores of Statements Posted at Q&A Sites. International Journal of Networked and Distributed Computing. 2022;10;1-2: 11–20. Available: https://link.springer.com/article/10.1007/s44227-022-00005-2
7. Yokoyama Y., Hochin T., Nomiya H. Using 4-gram to Obtain Factor Scores of Japanese Statements Posted at Q&A Sites. Proceedings of the 13th International Congress on Advanced Applied Informatics (AAI 2022-Winter). 2022: 25–31. Available: https://ieeexplore.ieee.org/document/10123522
8. Yokoyama Y. Application of 5-gram to Obtain Factor Scores of Japanese Q&A Statements. Proceedings of the 14th International Congress on Advanced Applied Informatics (AAI 2023). 2023: 69-75. Available: https://ieeexplore.ieee.org/document/10371593
9. 石田基広. R によるテキストマイニング入門 (第 2 版). 森北出版; 2017: 94.
10. 小林雄一郎. R によるやさしいテキストマイニング[活用事例編]. オーム社; 2018: 86-87.
11. The R Project for Statistical Computing. [cited 7 Feb 2024]. Available: https://www.r-project.org.