International Journal of Networked and Distributed Computing (2022) 10:11–20
https://doi.org/10.1007/s44227-022-00005-2

RESEARCH ARTICLE

Application of 2-gram and 3-gram to Obtain Factor Scores of Statements Posted at Q&A Sites

Yuya Yokoyama[1] · Teruhisa Hochin[2] · Hiroki Nomiya[2]

Received: 12 October 2021 / Accepted: 4 November 2022 / Published online: 18 November 2022

Abstract
With a view to solving the mismatches between the ideas of questioners and respondents of Question and Answer (Q&A) sites, impression evaluation experiments have resulted in obtaining nine factors of impressions. Then through multiple regression analysis factor scores have been estimated by utilizing the feature values of statements, such as syntactic information, etc. Those factor scores calculated were subsequently employed for inspecting their potential to detect respondents who are expected and likely to appropriately answer a newly posted question. Nevertheless, our method so far has largely depended on the syntactic information extracted through morphological analysis. Moreover, the number of explanatory variables utilized for obtaining factor scores has been appreciably extravagant and complex. Thus, instead of morphological analysis, 2-gram was applied to the explanatory variables to estimate factor scores. The analysis result with the application of 2-gram has led to greater estimation accuracy than the case of morphological analysis for all nine factors. For further perception and comparison, in this paper, 3-gram was applied to the feature values in place of 2-gram or morphological analysis, in a similar fashion as the previous analysis using 2-gram. Further analysis has shown that 2-gram and 3-gram outperform morphological analysis in terms of estimation accuracy. Comparing the results for the nine factors, 2-gram showed the best results. It could also be suggested that a mere 2-gram or 3-gram would be sufficient in applying N-gram as syntactic information of the feature values to estimate factor scores.

Keywords Q&A site · Factor score · Multiple regression analysis · 2-gram · 3-gram

1 Introduction

Recently, there have been increasing numbers of people utilizing Question and Answer (Q&A) sites, which are communities where users can manually post questions and answers, such as Yahoo! Chiebukuro (Y!C)[1] [1]. These Q&A sites are thought of as databases which encompass massive amounts of knowledge to resolve a variety of matters. The basic flow of a Q&A system is as follows: a user posts a question, and others might respond. The questioner chooses the most appropriate answer as the "Best Answer" (BA) and provides the respondent with awards as a token. The BA is the response the questioner subjectively finds most fulfilling.

With more users of Q&A sites registered and questions posted, it is getting more troublesome for respondents to pick up questions that coincide with their specialty and interests. Hence, a question given by a user might not be browsed and replied to by qualified respondents. In addition, though Q&A sites are becoming the collective knowledge for society, inappropriate answers can also be accumulated. Many an inappropriate answer statement could be posted as well. Thus, no appropriate respondents could result in mismatching and the following issues:

- Inappropriate answers may confuse the questioner and spread wrong knowledge.

✉ Yuya Yokoyama
y_yokoyama@mei.kpu.ac.jp

Teruhisa Hochin
hochin@kit.ac.jp

Hiroki Nomiya
nomiya@kit.ac.jp

[1] Graduate School of Life and Environmental Sciences, Kyoto Prefectural University, Kyoto, Japan

[2] Faculty of Information and Human Sciences, Kyoto Institute of Technology, Kyoto, Japan

[1] Y!C is a Japanese Q&A site that started their service in 2004.

- The shortage of necessary knowledge prevents respondents from providing proper answers, leaving the question unsolved.
- Abusive words, slander, or statements against public order and standards of decency might offend users.

Hence, requiring respondents to be users who are expected to provide appropriate answers is essential for storing appropriate answer statements. For the purpose of solving the issues explained above, a number of prior studies researching Q&A sites [2–9] with the employment of textual features or link analysis have been reported. Nevertheless, these works have yet to take into consideration the tendencies of the written styles of the users. Moreover, it is hard to say that a method to introduce appropriate respondents to a questioner has been settled yet. Thus, by gauging the impressions made by the statements, the objective of our work is to introduce appropriate respondents to a questioner. The promotion and extension of our work will contribute to the growing sphere of mere appropriate answer statements and make Q&A sites invaluable for societies, resulting in the swift and efficient promotion of social activities.

The aim of our work is thus to pose questions to users qualified to post proper answers to them, leading to curtailment of the problematic issues stated earlier. Through factor analysis applied to the experimental results, nine factors depicting the impression of Q&A statements have been captured [10]. Factor scores have then been estimated through multiple regression analysis from the 77 feature values of statements [11].

However, our method so far has largely depended on the syntactic information (Syn-Info) obtained through morphological analysis[2] (MA). In addition, the number of explanatory variables (EVs) is so enormous, resulting from regarding quadratic terms,[3] that the multiple regression equations to estimate factor scores employing them become tremendously complicated. Therefore, we have proceeded to estimating factor scores by utilizing the feature values of Syn-Info extracted through N-gram, one of the syntactic analysis methods like MA. [12]. As an initial step of N-gram, N was set to 2 in our previous work published as a conference proceeding [12].

In the previous analysis utilizing 2-gram instead of MA, in performing multiple regression analysis, the feature values based on 2-gram and those other than the Syn-Info were collectively employed as EVs, whereas the factor scores of their respective nine factors were set as respondent variables [12]. The analysis results have indicated that, for all these factors, the estimation result utilizing 2-gram has been nearly similar to or greater than the result employing MA [12]. Additionally, unlike the former method using MA where the quadratic term was indispensable for great estimation accuracy, a monadic term alone could be adequate for estimating factor scores and would contribute to fewer EVs with the simplification of the analysis results [12].

As an initial step to applying the N-gram so far, a mere 2-gram has been applied to the feature values of Syn-Info [12]. Therefore, in this paper, 3-gram is applied in place of 2-gram or MA. Similar to the previous analysis using 2-gram, through multiple regression analysis, the feature values based on 3-gram and those other than the Syn-Info are collectively utilized as EVs, whereas the factor scores are used as respondent variables. The further analysis results has shown that applying 2-gram and 3-gram show better estimation accuracy than MA. Comparing estimation accuracy for all nine factors, 2-gram shows the best results. It could also be suggested that in applying N-gram as Syn-Info, a mere 2-gram or 3-gram would be sufficient.

The rest of this paper is composed as follows. Section 2 introduces related works. As with our previous works, Sect. 3 summarizes obtaining factors of statement scores and estimating them. As with our previous work on applying N-gram, Sect. 4 explains multiple regression analysis utilizing 2-gram. Then Sect. 5 presents multiple regression analysis using 3-gram. Section 6 discusses considerations toward our analysis results are discussed. Finally, Sect. 7 concludes the paper.

2 Related Works

Numerous prior works investigating Q&A sites have been reported as follows: estimating BAs [2, 3]; introducing users to answer statements [4–6]; inspecting the quality or tendency of answer statements [7–9]; etc.

2.1 Estimation of BAs

Several works have tackled the estimation of BAs. Blooma et al. utilized a respective set of both five textual and non-textual features to predict the BAs [2]. Their analysis results have conveyed that textual features influenced the quality of the answers more than the non-textual ones did.

Calefato et al. assessed twenty-six BA prediction models in the following two steps [3]. Firstly, they studied the performance of models in predicting BAs in Stack Overflow[4] [13]. Then, they evaluated the performance in a

[2] Morphological analysis is known as one form of syntactic analysis. This method breaks down statements into their smallest morphemes to obtain their information such as frequency, Part-of-Speech, etc.

[3] A quadratic term is the product of an explanatory variable.

[4] Stack Overflow is one of the most popular Q&A sites for software engineers.

cross-platform setting where the prediction models were trained on Stack Overflow and tested on other Q&A sites. Their analysis results showed that the choice of the classifier and automated parameter tuning would play a significant role in predicting the BA. It has also been shown that their method of BA prediction issues is generalizable across technical Q&A sites.

2.2 Introducing Users to Answer Statements

Several research studies have proposed introducing users to answers. Zhang et al. tackled the issues where the patients' current usage of clinical data is considerably limited because of the technical nature of the clinical report [4]. With the rapid tendency for patients using online resources, e.g., Q&A sites, to acquire the knowledge by themselves, they analyzed Q&A statements posted in a Q&A site in order to shed light on what kinds of support people are providing to and receiving from the community and what contextual information they provide to deduce relevant answers. The analysis results revealed that users provided both objective and subjective information to the community. This emphasizes the importance of developing mechanisms to address the problem of the quality of online health information.

Haq et al. have done research on the Q&A site reputation through Quora, which is a Q&A platform that integrates elements of social networks to the traditional Q&A model [5]. In their recent study, they examined the impact of anonymity on the linguistic patterns, which were considered as playing a vital role in the involvement and grasp of the content. They then further developed their research on user interaction to demographics and analyzed its effect on topics engagement. They demonstrated that anonymity does not impact the polarity; and that anonymous answers and non-anonymous ones are drastically different from the viewpoints of length, subjectivity, and lexical diversity. It has been shown as well that stronger subjectivity contributes to more extreme polarity, partially because of the self-experience argued in the anonymous content.

Through a broad review of the present literature on expert recommendation, Yang et al. proposed four challenges [6]. Firstly, extant recommendation methods disregard the users' willingness to keep contributing within the online knowledge community. Another proposal is insufficient information in user profiles which hinders identifying potential experts. Thirdly, recommending experts as a collaborative group rather than looking for familiar individuals could drastically enhance the recommended answer rate. Finally, it is vital to regard the self-evolution of present expert recommendation approaches.

2.3 Quality or Tendency of Answer Statements

Several works have inspected the quality or tendency of answer statements. Bornfeld et al. explored the influence of vote and comment feedback mechanisms on the survival of answer providers after posting their first answer [7]. Their analysis results showed a strong correlation between votes and comments after the first post.

With a view to improving the professionalism of the social Q&A community and lead ordinary users to post high-quality answers, Shi et al. focused on the answer contents by disregarding the differences in the ability of respondents and evaluations of other users [8]. Through the relevant literature reviews, three dimensions were constructed: text features, rhetorical features, and emotional features of answer content. Nine features were then identified that might influence the quality of answer contents. Their analysis results have provided suggestions for users to post higher quality answers in terms of content for the functional optimization of social Q&A communities from the perspective of user requirements.

Li et al. have explored the characteristics of high-quality academic answer statements across different question types to facilitate the academic social Q&A sites to recommend high-quality answers to users on the basis of different question types [9]. Their analysis results have revealed that for discussion-seeking questions, users put more weight on the authority of respondents and whether the answer contains social elements, while for information-seeking questions, users focus more on whether the answer refers to the theoretical basis.

2.4 Summary

Although these prior works have primarily developed their research by employing textual features or link analysis, the tendency of answer statements have not been adequately taken into consideration. Some users may write in a polite style, while others might prefer to post their response in a ruder tone. Some commonly prefer abstract words, whereas others are apt to use more concrete ones. On the contrary, we focus on using impressions on top of textual features. In addition, despite several prior studies in the literature that introduce users to answer statements as described [4–6], a method to introduce appropriate respondents to a questioner has yet to be contrived. Therefore, using the impression of statements, our work aims to introduce appropriate respondents to a questioner.

3 Previous Works

3.1 Factors of Statements

To evaluate impressions of answer statements, an evaluation experiment was performed with the cooperation of 41 evaluators. They were asked to evaluate the style or content of statements and allocate five-level labels from a list of 50 impression words [10]. The experimental materials were 12 sets of Q&A statements composed of the respective three sets from four categories: Auction, PC, Love, and Politics & social issues. These materials were selected from those virtually posted at Y!C [2] in 2005 [10].

Factor analysis was then applied to the experimental results to obtain factors. The factors indicated the nature of a statement, as interpreted through the several impression words allotted to the statement. These factors were named *accuracy, displeasure, creativity, ease, persistence, ambiguity, moving, effort,* and *hotness*. The factor scores were also obtained to use in describing the characteristics of Q&A statements.

3.2 Estimation of Factor Scores

3.2.1 Feature Values of Statements

At this point, the factor scores were calculated for merely the sixty experimental materials utilized in the experiment explained in Sect. 3.1. With the aim of estimating the factor scores of any statements, multiple regression analysis was performed on their 77 feature values [11]. These feature values adopted are shown in Table Table 4 Feature values on 2-gram [12]gFeature values: Syn-Info (2-gram) g78[Noun-Part]g79[Part-Verb]g80[Part-Noun]g81[Noun-Noun]g82[Sign-Noun]g83[Verb-Aux]g84[Part-Sign] g85[Sign-Part]g86[Aux-Part]g87[Noun-Aux]g88[Aux-Sign]g89[Verb-Noun]g90[Noun–Verb]g91[Aux-Noun] g92[Aux-Aux]g93[Sign-Sign]g94[Part-Part]1. They are explained and summarized in the following five categories [11]:

(1) Syntactic Information (Syn-Info)
First, Syn-Info was utilized as the feature values of statements including statistics of statements, e.g., number/length of statements, and number/percentage of Part-of-Speeches (e.g., nouns, verbs etc.), etc. Specific marks such as exclamation and question marks were employed as well [11].

(2) Word Imageability (WI)
WI was also regarded as the feature values of statements [11]. WI is a subjective attribute implying how diverse imaginations can be recalled from words. The characteristic value of WI ranges from 1 to 7 [11].

(3) Closing Sentence Expressions (Closings)
Closings were included in the feature values as well [11]. The fundamental Japanese words adopted were "zo," "da," "yo," "ne," "ka," "na," "shi," "desu," "masu," "tai," and "nai" [11]. The feature values of Closings consist of the closing sentence words, the appearance, and closing sentence appearances as well as those words themselves. Here, the "closing word" indicates the appearance of the word at the end of a sentence.
Closing also includes the words "desuka," "naidesu," "masuka," and "mashita," which consist of two words of either "desu," "ka," "nai," "desu," and "masu."

(4) Word Familiarity (WF)
WF is an index indicating how familiar people feel or think either aurally or visually with a word [11]. The score of WF ranges from 1 to 7.

(5) Notation Validity (NV)
NV indicates the validity of a word and is evaluated by an index ranging from 1 to 5 [11]. A word can possess multiple different styles or meanings. Taking an example of the Japanese word "kosho," it could mean "breakdown," "lake," "name," etc., and written in the style of Chinese characters, hiragana or katakana characters, or their mixtures thereof.

3.2.2 Estimation Result

Multiple regression analysis was performed on the sixty Q&A statements utilized as the experimental materials in Sect. 3.1. Based on 77 monadic EVs, a total of 281 quadratic terms were set as explanatory variables, while factor scores for the nine factors were used as respondent variables.

The analysis result has shown that multiple correlation coefficients (MCCs), which indicate the estimation accuracy, were over 0.9 for all the nine factors [11]. Thus, all nine factors showed very good estimation accuracy.

4 Multiple Regression Analysis Using 2-gram

4.1 Aim

As summarized in Sect. 3.2, our method so far was largely dependent on the Syn-Info extracted through morphological analysis (MA). Moreover, employing quadratic terms has resulted in enormous EVs, leading to considerable complicated multiple regression equations utilized for estimating factor scores. Therefore, this paper aims to estimate factor

Table 1 77 Feature values of statements used for estimating factor scores [11]

(a) Syntactic information (Syn-Info)

g	Feature values (Syn-Info)	g	Feature values (Syn-Info)
g1	Auxiliary verbs (vocabulary)	g19	Full-size characters (%)
g2	Prefixes	g20	Alphanumeric characters (%)
g3	Signs (vocabulary)	g21	Full-size alphanumeric characters (%)
g4	Sentences	g22	Nouns (%)
g5	Average length of sentences (letters)	g23	Adjectives (%)
g6	Katakanas (word)	g24	Adverbs (%)
g7	Full-size characters (word)	g25	Pre-noun adjectivals (%)
g8	Full-size alphanumeric characters (word)	g26	Conjunctions (%)
g9	Adjectives (word)	g27	Interjections (%)
g10	Adverbs (word)	g28	Exclamation marks
g11	Pre-noun adjectivals (word)	g29	Question marks
g12	Conjunctions (word)	g30	Periods
g13	Interjections (word)	g31	Commas
g14	Hiraganas (%)	g32	Middle dots
g15	Chinese characters (%)	g33	Three dot leaders
g16	Katakanas (%)	g34	Quotation marks
g17	Signs (%)	g35	Parentheses
g18	TTR	g36	Slash characters

(b) Word imageability (WI)

g	Feature values: WI
g37	WI over 4.0 below 5.0 (word)
g38	WI over 6.5 below 7.0 (word)

(c) Closing sentence expressions (closing)

g	Feature values: closing	g	Feature values: closing
g39	"ka" (word)	g52	"zo" (%)
g40	"na" (word)	g53	"da" (%)
g41	"shi" (word)	g54	"yo" (%)
g42	"tai" (word)	g55	"ne" (%)
g43	"nai" (word)	g56	"ka" (%)
g44	"da" (cl-word)	g57	"desu" (%)
g45	"ka" (cl-word)	g58	"masu" (%)
g46	"na" (cl-word)	g59	"nai" (%)
g47	"shi" (cl-word)	g60	"ka" (closing (%))
g48	"desu" (cl-word)	g61	"desuka" (word)
g49	"masu" (cl-word)	g62	"naidesu" (word)
g50	"tai" (cl-word)	g63	"masuka" (word)
g51	"nai" (cl-word)	g64	"mashita" (word)

(d) Word familiarity (WF)

g	Feature values: WF
g65	WF percentage of words
g66	WF over 6.5 below 7.0 (vocabulary)
g67	WF over 4.0 below 5.0 (word)
g68	WF over 5.0 below 6.0 (word)
g69	WF over 5.5 below 6.0 (word)
g70	WF over 6.0 below 7.0 (word)

Table 1 (continued)

(d) Word familiarity (WF)	
g	Feature values: WF
g71	WF over 6.0 below 6.5 (word)
(e) Notation validity (NV)	
g	Feature values: NV
g72	NV percentage of words
g73	NV over 3.0 below 4.0 (word)
g74	NV over 3.5 below 4.0 (word)
g75	NV over 4.0 below 5.0 (word)
g76	NV over 4.0 below 4.5 (word)
g77	NV over 5.0 below 6.0 (word)

Table 2 The original Japanese statements of QA04 and their English translations [12]

QA04	Statements
Japanese (Original)	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。
English (Translation)	I am a beginner of using computers. I have printed out images I took with a digital camera, but they turned out dark. If anybody knows how to make them brighter, please answer my question.

scores employing the feature values of Syn-Info extracted through N-gram in place of MA. Using N-gram ought to result in higher estimation accuracy and provide more simplified equations to calculate factor scores.

4.2 N-gram

N-gram is also known as another method of syntactic analysis along with MA. N-gram depicts the adjacent sequence of *N* units of characters, morphemes, or Part-of-Speeches. Here, *N* is an arbitrary integer larger than 1 [14]. One question statement out of the sixty Q&A statements explained in Sect. 3.1 is utilized to show an N-gram Part-of-Speech example. The original Japanese question statements and their English translations are shown in Table 2. As a matter of convenience, the question is denoted as "QA04."

As for the 2-gram of QA04, their Part-of-Speeches, examples and frequencies are shown in Table 3. The column entitled "2-gram" have both literal notations and abbreviations. The notations "Noun," "Verb" and "Sign" are used as they are, whereas "Adjective," "Particle" and "Auxiliary" are abbreviated as "Adj," "Part" and "Aux," respectively. Thus, taking an example of the notation [Sign–Adj] shown in the first row, the 2-gram consists of a sign and an adjective. This provides one respective example each per 2-gram extracted from QA04 as shown in the column entitled "Example."

Table 3 2-Gram and frequency for QA04 [12]

2-Gram	Example	Frequency
[Sign-Adj]	[、-明るい]	1
[Sign-Noun]	[。-デジカメ]	1
[Adj-Verb]	[明るい-する]	1
[Adj-Noun]	[暗い-の]	1
[Part-Sign]	[が-、]	1
[Part-Adj]	[が-暗い]	1
[Part-Verb]	[で-撮る]	1
[Part-Noun]	[の-方]	4
[Aux-Sign]	[ます-。]	2
[Aux-Part]	[です-が]	1
[Aux-Noun]	[た-ところ]	2
[Verb-Aux]	[する-ます]	3
[Verb-Noun]	[する-方法]	1
[Noun-Part]	[画像-を]	6
[Noun-Aux]	[初心者-です]	2
[Noun–Verb]	[お願い-する]	2
[Noun-Noun]	[パソコン-初心者]	4

4.3 Analysis Method of 2-gram

In our previous analysis, 2-gram of Part-of-Speech was tentatively used instead of MA [12]. Here, 2-gram was applied to the sixty Q&A statements used for the experiment and stated in Sect. 3.1 to extract the feature values of 2-gram. Here, 2-gram was processed using R[5] [15]. At R, the library entitled RMeCab is installed so that N-gram as well as MA can be processed.

[5] R is a famous programming language and software environment to process statistics.

Table 4 Feature values on 2-gram [12]

g	Feature values: Syn-Info (2-gram)
g78	[Noun-Part]
g79	[Part-Verb]
g80	[Part-Noun]
g81	[Noun-Noun]
g82	[Sign-Noun]
g83	[Verb-Aux]
g84	[Part-Sign]
g85	[Sign-Part]
g86	[Aux-Part]
g87	[Noun-Aux]
g88	[Aux-Sign]
g89	[Verb-Noun]
g90	[Noun–Verb]
g91	[Aux-Noun]
g92	[Aux-Aux]
g93	[Sign-Sign]
g94	[Part-Part]

Similar to the analysis stated in Sect. 3.2, multiple regression analysis was run to obtain factor scores of the nine factors, which were used as the respondent variable. Meanwhile, as for a part of EVs, through trial and error, seventeen 2-gram of Part-of-Speeches were employed as feature values of Syn-Info, which are denoted as g78–g94 as summarized in Table 4. In this analysis, the feature values of Syn-Info on the basis of 2-gram (g78–g94) were used in place of those based on MA (g1–g36). In conjunction with WI, Closings, WF, and NV (g37–g77), a total of 68 feature values (g37–g94) were employed as EVs.

4.4 Estimation Result

EVs with absolute values of the standardized partial regression coefficient (SPRCs) bigger than 0.1 were focused on. The EVs are summarized in Table 5. Among the EVs satisfying the condition of SPRCs over 1.0, the maximum three positive/negative strongest EVs were extracted for each factor. In the column entitled "FV," the classifications of feature values are shown that coincide with the column entitled "EV" and that are shown in Tables 1 and 4.

Table 5 Explanatory variable (EV) and feature value (FV) with higher standardized partial regression coefficient (SPRC): 2-gram [12]

1st (Accuracy)			2nd (Displeasure)			3rd (Creativity)		
EV	FV	SPRC	EV	FV	SPRC	EV	FV	SPRC
g84	2-gram	1.27	g79	2-gram	5.65	g65	WF	3.68
g87	2-gram	1.24	g70	WF	3.28	g39	Closing	3.01
			g86	2-gram	2.05	g85	2-gram	2.83
			g39	Closing	−2.60	g72	NV	−2.18
			g87	2-gram	−2.86	g70	WF	−3.49
			g83	2-gram	−3.09	g79	2-gram	−6.90
4th (Ease)			5th (Persistence)			6th (Ambiguity)		
EV	FV	SPRC	EV	FV	SPRC	EV	FV	SPRC
g80	2-gram	3.71	g80	2-gram	3.00	g78	2-gram	1.66
g79	2-gram	2.82	g83	2-gram	1.59	g86	2-gram	1.29
g90	2-gram	−1.17	g45	Closing	1.35	g79	2-gram	−1.89
g78	2-gram	−5.30	g86	2-gram	−1.23	g80	2-gram	−2.32
			g78	2-gram	−2.40			
7th (Moving)			8th (Effort)			9th (Hotness)		
EV	FV	SPRC	EV	FV	SPRC	EV	FV	SPRC
g78	2-gram	4.12	g83	2-gram	2.85	g87	2-gram	5.21
g87	2-gram	2.93	g45	Closing	2.25	g83	2-gram	4.54
g85	2-gram	2.19	g78	2-gram	2.20	g85	2-gram	3.54
g80	2-gram	−2.43	g65	WF	−1.41	g90	2-gram	−2.98
g86	2-gram	−3.17	g79	2-gram	−1.61	g70	WF	−5.90
g79	2-gram	−5.20	g37	WI	−1.71	g79	2-gram	−12.08

Table 6 3-gram and frequency for QA04

g	Feature values: Syn-Info (3-gram)
g95	[Sign-Noun-Part]
g96	[Noun-Noun-Noun]
g97	[Part-Sign-Noun]
g98	[Part-Verb-Noun]
g99	[Sign-Noun-Noun]
g100	[Verb-Noun-Part]
g101	[Noun-Aux-Part]
g102	[Verb-Aux-Noun]
g103	[Noun-Part-Adj]
g104	[Aux-Aux-Sign]
g105	[Sign-Sign-Sign]
g106	[Part-Sign-Adv]
g107	[Noun-Verb-Aux]
g108	[Aux-Noun-Aux]
g109	[Sign-Noun-Sign]
g110	[Noun-Verb-Noun]
g111	[Noun-Noun-Aux]

Similar to the results with MA depicted in Sect. 3.2, MCCs outnumbered 0.9 for all the nine factors. Moreover, MCCs were improved with the application of 2-gram rather than MA. Therefore, estimation accuracy utilizing 2-gram showed results almost equivalent or superior to those employing MA.

5 Multiple Regression Analysis Using 3-gram

5.1 Aim

In applying N-gram to our method so far, a mere 2-gram was applied to the feature values of Syn-Info. For further analysis, a bigger unit of N-gram than 2-gram must also be applied and analyzed. Therefore, in this paper, 3-gram was applied instead of 2-gram or MA. Similar to the previous analysis using 2-gram, the analysis method using 3-gram was performed through multiple regression analysis. The analysis result using 3-gram was then compared with those using 2-gram or MA to validate the effectiveness of the application of 3-gram.

The feature values based on 3-gram and those besides the Syn-Info were collectively utilized as EVs, whereas the factor scores for the respective nine factors were employed as respondent variables.

Table 7 Explanatory variable (EV) and feature value (FV) with higher standardized partial regression coefficient (SPRC): 3-gram

1st (Accuracy)			2nd (Displeasure)			3rd (Creativity)		
EV	FV	SPRC	EV	FV	SPRC	EV	FV	SPRC
g70	WF	1.90	g76	NV	1.16	g64	Closing	2.83
g37	WI	1.47	g73	NV	−0.77	g100	3-gram	2.03
g43	Closing	1.14				g65	WF	1.95
g62	Closing	−1.06				g76	NV	−1.68
g100	3-gram	−1.54				g37	WI	−2.09
g64	Closing	−1.65				g70	WF	−2.36
4th (Ease)			5th (Persistence)			6th (Ambiguity)		
EV	FV	SPRC	EV	FV	SPRC	EV	FV	SPRC
g65	WF	1.42	g45	Closing	0.97	g66	WF	1.04
g44	Closing	1.29	g60	Closing	−0.47	g98	3-gram	−1.22
g70	WF	−1.10				g43	Closing	−1.31
g76	NV	−1.34				g70	WF	−1.34
g72	NV	−1.72						
7th (Moving)			8th (Effort)			9th (Hotness)		
EV	FV	SPRC	EV	FV	SPRC	EV	FV	SPRC
g100	3-gram	1.56	g100	3-gram	2.08	g66	WF	2.04
g98	3-gram	−1.30	g68	WF	1.61	g65	WF	1.87
			g59	Closing	1.59	g73	NV	1.77
			g107	3-gram	−1.37	g76	NV	−2.04
			g43	Closing	−1.58	g98	3-gram	−2.11
			g37	WI	−1.61	g70	WF	−3.26

Table 8 Comparison of multiple correlation coefficients (MCCs)

Factor	MA	2-gram	3-gram
1st (Accuracy)	1.00	0.989	0.993
2nd (Displeasure	0.947	0.999	0.987
3rd (Creativity)	0.877	0.981	0.998
4th (Ease)	0.908	0.990	0.995
5th (Persistence)	0.966	0.993	0.976
6th (Ambiguity)	0.899	0.998	0.994
7th (Moving)	0.997	0.999	0.996
8th (Effort)	0.904	0.995	0.968
9th (Hotness)	0.954	0.995	0.998

5.2 Analysis Method of 3-gram

Similar to the analysis method using 2-gram stated in Sect. 4.3, multiple regression analysis was processed. Factor scores of the nine factors are utilized as the respondent variable. In order to easily and directly compare the analyses between 2 and 3-gram, as for the feature values of Syn-Info, the amount of 3-gram extracted is the same as that of 2-gram, seventeen. These feature values are denoted as g95–g111 shown in Table 6.

In this analysis, the feature values of Syn-Info on the basis of MA (g1–g36) are replaced by those based on 3-gram (g95–g111). In conjunction with WI, Closings, WF, and NV (g37–g77), a total of 68 feature values (g37–g94) are utilized as EVs. Most of the abbreviations are already explained in Sect. 4.2, except one that had not appeared in Tables 3 or 4: "Adv" stands for adverb and is extracted as one component of 3-gram [Part-Sign-Adv] (g106).

5.3 Estimation Result

Similar to the former method utilizing 2-gram stated in Sect. 4.2, EVs with absolute values of SPRCs larger than 1.0 are summarized in Table 7. Among the EVs meeting the condition of SPRC over 1.0, the maximum three positive/negative strongest EVs are shown for each factor. However, there are several cases where the absolute values of SPRCs are below 1.0 for all the EVs: negative SPRC for the 2nd factor and both positive/negative SPRCs for the 5th factor. For these cases, only the one largest EV for positive/negative SPRC is shown. The columns entitled "FV" is explained in Sect. 4.4. The analysis result thus conveys that the MCCs for all nine factors outscore 0.9.

6 Considerations

In order to compare MCCs among 3-gram, 2-gram, and MA, these results are summarized in Table 8. From the viewpoints of MCCs, the figures with the case using 2-gram show the best results for five factors (2nd, 5th, 6th, 7th, and 8th), followed by those with the case employing 3-gram for three (3rd, 4th, and 9th) and those utilizing MA for one (1st). From these comparisons, using 2-gram is best among these three cases. Nevertheless, as a whole, MCCs are improved with the application of 2-gram or 3-gram. Therefore, it could be suggested that considering N-gram would outperform the analysis results using mere MA. It could also be suggested that considering 2-gram or 3-gram would be sufficient in applying N-gram. In other words, it would be unnecessary to analyze beyond 4-g with this method.

These results could result from regarding 2-gram or 3-gram, which convey the collocations among two/three words. However, the associations among words are disregarded with the cases of MA. From these standpoints, regarding 2-gram or 3-gram could be more productive for estimating factor scores of Q&A statements.

In addition, in the previous analysis using MA, quadratic terms were required for good estimation accuracy. With the usage of N-gram, by contrast, monadic terms alone would be adequate in estimating factor scores. Thus, N-gram contributes to limiting the process to much fewer EVs, which results in the simplification of multiple regression equations to obtain factor scores.

Nevertheless, the meanings or contents of Q&A statements have not been considered for our analysis so far. Hence, with a view to regarding them, a meaning analysis needs to be applied to our method in the future. Moreover, it is indispensable to investigate if our proposed method utilizing N-gram can be extended to other languages.

7 Conclusions

In this paper, 3-gram was applied instead of 2-gram or MA. Similar to our previous analysis using 2-gram, through performing multiple regression analysis, the feature values based on 3-gram, and those other than syntactic information were collectively utilized as explanatory variables, while the factor scores for the respective nine factors were set as respondent variables. As a result of this further analysis, in comparing estimation accuracy for the nine factors among the cases using 2-gram, 3-gram, and MA, 2-gram showed the best results. As a whole, applying 2-gram or 3-gram would improve estimation accuracy more than MA would. In addition, it could also be suggested that a mere 2-gram or 3-gram would be sufficient in applying N-gram as syntactic information to our method.

For future work, the meanings and contents of Q&A statements must be taken into account for the analysis. Moreover, with the feature values of syntactic information based on MA, the factor scores obtained were subsequently employed

for inspecting the possibility of detecting respondents who could be expected to post the appropriate answer to a newly posted question [16]. Therefore, whether the feature values based on 2-gram could be effective in finding appropriate respondents should be inspected and compared with the case of MA. Because most of the feature values used in this study are based on Japanese language materials, the generalization of these findings to other languages has to be addressed as another topic in our future work as well.

Acknowledgements This research was partially supported by the Japan Society for the Promotion of Science, Grant Numbers 26008587, 2015-2016, and Grant-in-Aid for Young Scientists, Grant Numbers 20K19933, 2020-2023.

Author Contributions YY data curation, formal analysis, funding acquisition, project administration, and writing the manuscript (original draft). TH and HN supervised the project and writing the manuscript (review and editing).

Data Availability Statement The data that support the findings of this study are openly available in Yahoo! Chiebukuro data provided by National Institute of Informatics at https://www.nii.ac.jp/dsc/idr/en/yahoo/chiebkr2/Y_chiebukuro.html, reference number [17].

Declarations

Conflict of Interest The authors declare they have no conflicts of interest.

References

1. Yahoo! Chiebukuro (URL, in Japanese), http://chiebukuro.yahoo.co.jp/, 2021–12–16
2. Blooma MJ, Chua AYK and Goh DHL (2008) A predictive framework for retrieving the best answer. In the Proceedings of 2008 ACM Symposium on Applied Computing (SAC08), pp, 1107–1111. https://doi.org/10.1145/1363686.1363944
3. Calefato F, Lanubile F, Novielli N (2019) An empirical assessment of best-answer prediction models in technical Q&A sites. Empir Softw Eng 24:854–901. https://doi.org/10.1007/s10664-018-9642-5
4. Zhang Z, Lu Y, Wilson C and He Z (2019) Making sense of clinical laboratory results: an analysis of questions and replies in a social Q&A community. In the Proceedings of the 17th World Congress on Medical and Health Informatics (MEDINFO 2019), pp 2009–2010. https://doi.org/10.3233/SHTI190759
5. Haq EU, Braud T and Hui P (2020) Community matters more than anonymity: analysis of user interactions on the Quora Q&A platform. In the Proceedings of the International conference series on Advances in Social Network Analysis and Mining (ASONAM 2020), pp 94–98
6. Yang Z, Liu Q, Sun B, Zhao X (2019) Expert recommendation in community question answering: a review and future direction. Int J Crowd Sci 3(3):348–372. https://doi.org/10.1108/IJCS-03-2019-0011
7. Bornfeld B and Rafaeli S (2019) When interaction is valuable: feedback, churn and survival on community question and answer sites: the case of stack exchange. In the Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS 2019), pp 789–799. https://doi.org/10.24251/HICSS.2019.096
8. Shi J, Shen H and Ma Q (2019) What kind of answer will be better: exploring the features of high-quality answer contents in social Q&A community. In the Proceedings of the 19th International Conference on Electronic Business (ICEB19), pp 558–562
9. Li L, He D and Zhang C (2019) Characterizing high-quality answers for different question types on academic social Q&A site. In the Proceedings of the 17th International Conference on Scientometrics and Informetrics (ISSI 2019), pp 2670–2671
10. Yokoyama Y, Hochin T, Nomiya H, Satoh T (2012) Obtaining Factors Describing Impression of Questions and Answers and Estimation of their Scores from Feature Values of Statements. Softw Netw Eng 413:1–13. https://doi.org/10.1007/978-3-642-28670-4_1 (**Springer**)
11. Yokoyama Y, Hochin T, Nomiya H (2014) Using feature values of statements to improve the estimation accuracy of factor scores of impressions of question and answer statements. Int J Affect Eng 13(1):19–26. https://doi.org/10.5057/ijae.13.19
12. Yokoyama Y, Hochin T, Nomiya H (2021) Application of 2-gram to obtain factor scores of statements posted at Q&A sites. In: Proceedings of the 8th ACIS international virtual conference on applied computing & information technology (ACIT 2021), pp 111–117. https://doi.org/10.1145/3468081.3471132
13. Stack Overflow (URL), https://stackoverflow.com, 2021–12–16
14. Ishida M (2017) Text Mining Introduction Using R (in Japanese), 2nd edn. Morikita Publishing, pp 94–99 (**ISBN978-4-627-84842-9**)
15. The R Project for Statistical Computing (URL), https://www.r-project.org, 2021–12–16
16. Yokoyama Y, Hochin T, Nomiya H (2019) Quantitative Evaluation of Potential Tendency Differences between English and Japanese in Detecting Appropriate Respondents at Q&A Sites. Int J Affect Eng 18(3):145–154. https://doi.org/10.5057/ijae.IJAE-D-18-00023
17. Distribution of "Yahoo! Chiebukuro data (2nd edition)" (URL), https://www.nii.ac.jp/dsc/idr/en/yahoo/chiebkr2/Y_chiebukuro.html, 2021–12–16.