Exploratory analysis of stylistic characteristics in Japanese Q&A communities

Takafumi Suzuki [1], Shuntaro Kawamura [2], Akiko Aizawa [1]

[1] National Institute of Informatics - 2-1-2 Hitotsubashi - Chiyoda-ku - Tokyo - Japan

[2] University of Tokyo - 7-3-1 - Bunkyo-ku - Hongo - Tokyo – Japan

Abstract

This study is an exploratory analysis of the stylistic characteristics of text submitted to Japanese Q&A communities. Along with the development of social media, Q&A communities are attracting much scholarly attention as important resources for analyzing online communication. In Q&A communities, people freely submit questions and answers; questions are classified into subject categories; and the best answers are selected. In this study, we analyze the stylistic characteristics of three types of submission, i.e., questions, best answers, and normal answers, in two different subject categories, i.e., 'personal computers and related devices' and 'love and human relations advice'. The results show that the textual styles clearly distinguished these six classes of text and clarified their respective characteristics. Our findings provide useful knowledge about how people differ in their communication styles regarding subject categories and on how people select communication styles. This study will contribute to research into discovering current online communication styles.

Keywords: computational stylistics, online communication styles, Japanese Q&A communities

1. Introduction

Through the development of the Web, various new texts media have appeared (Aitchison and Lewis, 2003). In particular, text in social media such as Wikis, blogs, and SNS are produced by the users themselves, reflect their users' interests, and reveal new styles of online communication. The textual characteristics of such media should be useful for tracking changes in language usage and current communication styles on the Web, especially in the context of Japanese.

Among the many social media, Q&A communities where people freely submit questions and answers online are attracting much scholarly attention. In Q&A communities, questions are classified into subject categories, and the best answers are selected by some criteria [1]. Thus, the text submissions provide us with fruitful examples of how people differ in their communication styles regarding subject categories and of how people select communication styles to fit their circumstances.

To analyze the text submissions, we focus on their styles. Style, i.e., textual characteristics independent on the content of the text, is 'how it is mentioned in the text' (Argamon et al., 2007), and knowledge of style has various new applications, such as authorship profiling, sentiment analysis, and computational sociolinguistics, as well as conventional applications, such as authorship attribution and genre discrimination (Argamon et al., 2007; Koppel et al., 2009; Stamatatos, 2009; Suzuki, 2009). Styles are useful for determining, for example, the

[1] Regarding the data we used in this study, all the best answers were selected by questionnaire.

author's personality, feelings, sentiments; thus, we thought they would be good for analyzing communication styles of Q&A communities.

This study constitutes an exploratory analysis of the stylistic characteristics of texts submitted to Japanese Q&A communities. We compare two different types of submission, factual questions [2] and personal advice questions [3] (Harper et al., 2008; Miura and Kawamura, 2008), which are typical, yet completely different types of questions. By analyzing the stylistic characteristics of questions, best answers, and normal answers, of these two categories, we can see how people vary their communication styles across subject categories and how people select communication styles. Note that we conducted only an exploratory analysis, because there are few studies focusing on the textual characteristics of Q&A communities, especially in Japanese. By analyzing the styles of the submissions, we tried to derive knowledge on how people communicate with each other online; such knowledge is becoming more and more important since people are spending more time in virtual spaces. Our study also provides fundamental knowledge for many IR and NLP tasks, *e.g.*, good answer estimation, automatic paraphrasing, and automatic conversation generation.

2. Data

We used Yahoo! Chiebukuro (Japanese version of Yahoo! answers) data provided to National Institute of Informatics by Yahoo Japan Corporation. This data includes 3,116,009 questions, 3.116.008 best answers, 10.361.777 normal answers that were submitted during the period from April 2004 to October 2005. All the submissions are classified into subject categories.

We selected two categories for our analyses, 'personal computers and peripheral devices' (PC), and 'love and human relationships advice' (LH). The former category is a typical one that includes factual questions, whereas the latter category is a typical one that includes personal advice questions. We collected texts of questions (Q), best answers (BA), normal answers (NA), per month [4] and applied morphological analysis using MeCab [5], a Japanese morphological analysis system. We assigned parts-of-speech tags by using MeCab and calculated the number of tokens per submission and frequencies of function words per text.

As features, we used the bag-of-words of the relative frequencies of function words, i.e., functional nouns (noun-dependent and noun-pronominal), adnominals, conjunctions, particles and auxiliary verbs. As function words independent of the content represent the *affect*, *genre*, *register* and *personality* of the texts (Argamon et al., 2007), and are effective for sociolinguistic analysis as well as stylistic text classification (Garcia and Martin, 2007; Grieve, 2007; Suzuki, 2009), they are appropriate features for our purpose. In Japanese, particles and auxiliary verbs are strongly related to the modality of the text (Otsuka et al., 2007) and adnominals, conjunctions, and some particles represent the logicality and readability of the text (c.f., Otsuka et al., 2007; Tuldava, 1993), while some functional nouns can represent explanation patterns.

It is better for our purpose to use deeper-order part-of-speech tags of particles [6] and the stemming version of auxiliary verbs as they facilitate more meaningful interpretations [7]. Tab. 1 lists the

[2] They are, in other words, questions that have certain answers.
[3] They are, in other words, questions that have no certain answers.
[4] Text is usually analyzed on a per submission basis for many NLP and IR tasks, but our purpose here is to clarify the basic stylistic characteristics of six categories; thus, it is better to use the texts per month.
[5] mecab.sourceforge.net.
[6] 'Case particles' or 'conjunctive particles', etc.
[7] A particle can have different meanings when it is used in different second-order parts-of-speech, while the

number of submissions and the total number of tokens and types of function words in the six categories, while Tab. 2 lists the respective number of tokens and types for each part-of-speech.

			length of postings		
		number of postings	*mean*	*s.d.*	*c.v.*
PC	Q	171,867	52.69	34.68	65.82
	BA	171,848	55.93	43.36	77.53
	NA	302,839	37.59	31.85	84.72
LH	Q	210,124	70.79	53.03	74.90
	BA	210,105	68.11	48.72	71.54
	NA	1,206,457	45.57	37.83	83.00

Table 1: Basic data of our corpora 1

		functional nouns		*conjunctions*		*adnominals*	
		N	*V(N)*	*N*	*V(N)*	*N*	*V(N)*
PC	Q	509,223	192	40,422	120	62,992	63
	BA	378,526	205	52,669	132	64,197	75
	NA	490,155	211	57,245	136	75,677	75
LH	Q	1,119,160	213	89,206	133	163,096	78
	BA	1,038,136	219	96,351	137	143,042	86
	NA	3,901,410	225	345,121	146	538,682	94
all		7,436,610	227	681,014	149	1,047,686	97

		particles		*aux. Verbs*		*functional words*	
		N	*V(N)*	*N*	*V(N)*	*N*	*V(N)*
PC	Q	2,595,688	171	1,236,572	107	4,444,897	606
	BA	2,839,645	178	975,193	120	4,310,230	651
	NA	3,346,199	180	1,264,017	130	5,233,293	663
LH	Q	4,623,131	174	2,143,677	132	8,138,270	660
	BA	4,652,724	181	1,958,581	141	7,888,834	686
	NA	17,407,519	182	7,644,807	152	29,837,539	710
all		35,464,906	190	15,222,847	155	59,853,063	726

Table 2: Basic data of our corpora 2

3. Methods

After we observe the basic characteristics, we make a text-feature matrix, whose rows represent the texts per month and columns represent features (relative frequencies of each function word to the sum of all the function words). Then we apply principal component analysis and random forests.

First, we apply principal component analysis with the covariance matrix of the features. Principal component analysis enables us to view as a scatter plot and to clarify the factors classifying the texts as the principal component [8].

different forms of an auxiliary verb have the same meaning in Japanese.

[8] There are other methods of exploratory data analyses, *e.g.*, factor analysis, correspondence analysis, or multidimensional scaling. Even though there are no special rules for deciding which method is the best for

Next we apply random forests proposed by Breiman (2001). We replicated the original data matrix $M_{i,j}$ 1000 times with replacement, and extracted random subsets of variables from each replicated data. We constructed an unpruned decision tree for each sample by using the Gini index. We constructed a new classifier by conducting a majority vote of the set of trees. Two-thirds of the bootstrap samples were used for constructing the model and the other third were left for testing the model (out-of-bag test).

We calculated the variable importance using the following formula (Breiman, 2001):

$$VI_{acu} = \frac{mean(C_{oob} - C_{per})}{s.e.}$$

C_{oob}: number of votes cast for the correct class in the out-of-bag data

C_{per}: number of votes cast for the correct class when m variables are randomly permuted in the out-of-bag data

s.e.: standard error.

The mean value of subtractions for all trees formulated above represents the variable importance for a permuted variable. It represents the degree to which a class loses its specific character when one type of morpheme changes into another type of morpheme. This method calculates important variables directly contributing to the classification in the experiment; thus, it suits for our purpose best (Suzuki, 2009).

We used the macro average of F_1 values for evaluating the results. Random forests uses random numbers in the experiments; thus, we performed the experiments 100 times and calculated the mean F_1 values for these 100 experiments (Jin and Murakami, 2007).

4. Results and discussion

4.1. Basic Observation

Tab. 1 lists the results of the basic observation, number of submissions, and number of tokens for a submission (mean, standard deviations, coefficient of variation), for six categories. The results show that on average LH has a larger number of submissions and is longer in length than PC; BA is as long as Q; NA is shorter than others, and NA has larger variances than others.

Tab. 2 lists the respective numbers of tokens N and types ($V(N)$) for each part-of-speech and all the function words in six categories. These results show that BA has a larger N than Q in terms of conjunctions and particles and a smaller N in functional nouns and auxiliary verbs. The tendency of N is different between PC and LH adnominals. NA has larger N and $V(N)$ compared with Q and BA. These basic characteristics show the differences between PC/LH and Q/BA/NA, which we will discuss in Section 4.3 in detail.

4.2. Principal component analysis

We carried out a principal component analysis using the covariance matrix of the features [9]. Fig. 1 represents the scatter plot showing the first two principal components [10]. The proportion

respective data, we selected PCA because its results tend to be unambiguous, and thus it should be applied first (Jin, 2007).

[9] We also carried out the method using the correlation coefficient matrix, and found no significant differences between results.

[10] Texts are indicated by combinations of subject categories (PC/LH) and types (Q/BA/NA).

of variance accounted for by the first principal component was 56.17%, and the cumulative proportion of variances accounted for by the first two principal components was 93.37%. The first principal component mainly represents the axis of subject categories, because the texts in LH fell to the left side of the scatter plot and those in PC to the right side. The second principal component mainly represents the axis of questions/answers, because questions fell on the upper side, and answers on the lower side. Best answers and normal answers were clearly distinguished regarding PC, but not regarding LH. These results show that the difference between questions and answers is rather large, and there is still a difference between best and normal answers regarding writing styles. Three texts in a rather isolated position in the scatter plot (Q_LH04, Q_PC04, B_LH04, N_LH04) were submissions in April 2004, in which the total number of tokens (*N*) and number of submissions were small.

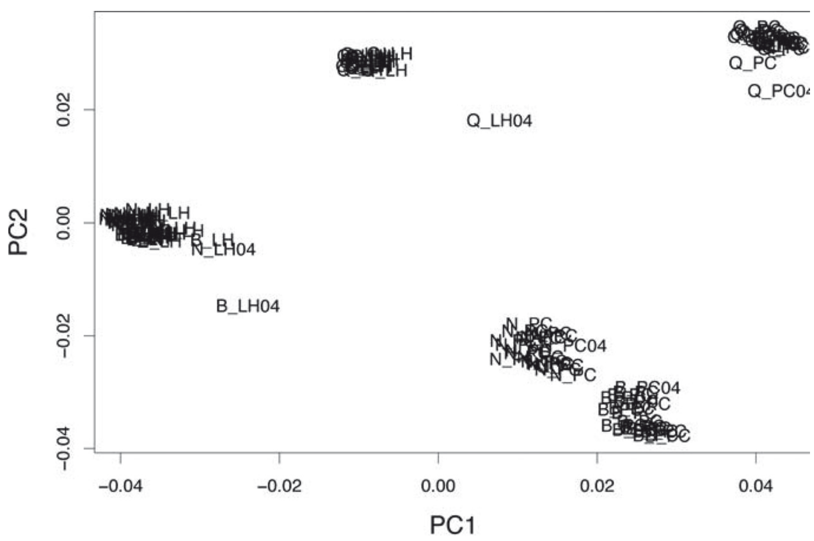

Figure 1: Scatter plot showing first two principal components

4.3. Feature selection by machine learning

We next carried out six-class classification experiments (Q_LH, Q_PC, B_LH, B_PC, N_LH, N_PC) by using random forests using the text-feature matrix of function words. Precision, recall rate, and F_1 value in the experiment was 97.06, 96.80, and 96.79 respectively.

Tab. 3 show the top 20 variables in the classification experiments using all the function words with their parts-of-speech, variable importance (VI_{acu}), and the notation that these variables were frequently used (H) or infrequently used (L) in the category in comparison with mean of other categories. These variables significantly contributed to the classification, and, thus, they represent class-specific function words. We shall discuss the results by conducting qualitative analyses. In this way, we will find many interesting issues that should be subjects of further, hypothesis-deductive research.

4.4. Comparison of subject categories regarding questions

Both categories include the particle 'ka' as a frequent expression. This is a typical expression for making questions and a salient character of the question type. LH includes many pronouns like 'watashi' (I; rank1), 'kare' (he; 2), 'boku' (I; 11), and 'kanojo' (she; 16), and adnominal 'konna' (such; 6) and noun-affix 'mitai' (like; 5) as frequent expressions; these rarely appeared in the PC category. These results represent the characteristic of the LH category wherein people tend to ask questions after they explain episode about themselves. The frequent expressions in the PC category includes the particle-conjunctive 'ga' (-(subjective); 7) and noun-affix 'no' (of; 4). These results represent the characteristic that people simply ask about people what they would like to know.

	Q_LH			Q_PC			B_LH		
rank	morphemes	VI_{acu}	H/L	morphemes	VI_{acu}	H/L	morphemes	VI_{acu}	H/L
1	watashi-noun-pronoun	2.142	H	da-aux. verb	1.982	L	mo-particle-dependency	2.728	H
2	kare-noun-pronoun	2.124	H	shi-particle-conjunctive	1.922	L	soshite-conjunction	2.622	H
3	ba-particle-conjunctive	2.123	L	nante-particle-adverbial	1.908	L	ni-particle-adverbializer	2.548	H
4	niyotte-particle-case	2.060	L	no-noun-affix	1.890	H	ka-particle-adverbial/coordinate/final	2.428	L
5	mitai-noun-affix	2.058	H	nai-aux. verb	1.879	L	desu-aux. verb	2.305	L
6	konna-adnomial	2.040	H	ja-particle-adverbial	1.869	L	tari-particle-coordinate	2.275	H
7	bakari-particle-adverbial	1.814	H	ga-particle-conjunctive	1.863	H	te-particle-conjunctive	2.145	L
8	ga-particle-conjunctive	1.763	H	ka-particle-adverbial/coordinate/final	1.861	H	n-aux. verb	2.078	L
9	nante-particle-adverbial	1.757	L	kamo-particle-adverbial	1.838	L	ja-particle-adverbial	2.008	H
10	ni-particle-adverbializer	1.728	L	yo-particle-final	1.815	L	nu-aux. verb	2.007	H
11	boku-noun-pronoun	1.720	H	dono-adnomial	1.805	H	to-particle-conjunctive	1.965	L
12	kamo-particle-adverbial	1.720	L	mo-particle-dependency	1.796	L	soretomo-conjunction	1.934	L
13	ja-particle-adverbial	1.719	L	desu-aux. verb	1.727	H	ga-particle-conjunctive	1.919	L
14	ka-particle-adverbial/coordinate/final	1.709	H	ha-particle-dependency	1.711	L	da-aux. verb	1.826	H
15	sore-noun-pronoun	1.702	L	demo-particle-adverbial	1.710	L	dono-adnomial	1.783	L
16	kanojo-noun-pronoun	1.675	H	na-particle-final	1.699	L	shi-particle-conjunctive	1.782	H
17	ta-aux. verb	1.671	H	tada-conjunction	1.693	L	sore-noun-pronoun	1.779	H
18	no-noun-affix	1.663	H	nitsuite-particle-case	1.675	H	anata-noun-pronoun	1.758	H
19	ne-particle-final	1.661	L	chinamini-conjunction	1.673	H	ni-particle-case	1.742	L
20	mo-particle-dependency	1.653	L	no-particle-case	1.670	L	kamo-particle-adverbial	1.736	H

	B_PC			N_LH			N_PC		
rank	morphemes	VI_{acu}	H/L	morphemes	VI_{acu}	H/L	morphemes	VI_{acu}	H/L
1	tadashi-conjunction	2.790	H	te-particle-conjunctive	2.918	L	shi-particle-conjunctive	2.301	H
2	no-particle-adnominalizer	2.782	H	nante-particle-adverbial	2.383	H	da-aux. verb	2.296	H
3	wo-particle-case	2.706	H	ja-particle-adverbial	2.308	H	nai-aux. verb	2.255	H
4	ni-particle-case	2.434	H	ni-particle-case	2.279	L	nante-particle-adverbial	2.173	H
5	no-noun-affix	2.369	L	to-particle-conjunctive	2.213	L	no-particle-adnominalizer	2.132	L
6	shi-particle-conjunctive	2.334	L	nai-aux. verb	2.181	H	no-noun-affix	2.005	L
7	da-aux. verb	2.291	L	ka-particle-adverbial/coordinate/final	2.126	L	ne-particle-final	2.005	H
8	n-noun-affix	2.286	L	mata-conjunction	2.119	L	ja-particle-adverbial	1.980	H
9	tte-particle-case	2.221	L	tari-particle-coordinate	1.980	H	ka-particle-adverbial/coordinate/final	1.964	L
10	nante-particle-adverbial	2.155	L	sonna-adnomial	1.869	H	tadashi-conjunction	1.947	L
11	nai-aux. verb	2.132	L	kara-particle-case	1.869	L	wo-particle-case	1.890	L
12	kurai-particle-adverbial	2.077	L	desu-aux. verb	1.861	L	kono-adnomial	1.866	L
13	ika-noun-affix	2.027	H	da-aux. verb	1.843	H	tai-aux. verb	1.795	H
14	ka-particle-adverbial/coordinate/final	1.936	L	anta-noun-pronoun	1.777	H	yo-particle-final	1.766	H
15	nani-noun-pronoun	1.914	L	wo-particle-case	1.735	L	n-noun-affix	1.761	H
16	tame-noun-affix	1.903	H	ni-particle-adverbializer	1.732	H	sonna-adnomial	1.734	H
17	ja-particle-adverbial	1.893	L	you-noun-affix	1.724	L	nu-aux. verb	1.731	H
18	ne-particle-final	1.833	L	sore-noun-pronoun	1.720	H	ga-particle-conjunctive	1.684	L
19	dare-noun-pronoun	1.826	L	dono-adnomial	1.694	L	mo-particle-dependency	1.682	H
20	beshi-aux. verb	1.820	L	ga-particle-conjunctive	1.681	L	それ-noun-pronoun	1.669	H

Table 3: Top twenty variables with high variable importance (VI_{aci}) for respective classes

4.5. Comparison of best and normal answers

Both answers in the LH category include particle-conjunctive 'tari' (- (paralel), 6) as a frequent expression, while it does not appear in the PC answers. 'tari' is a parallellization expression used as '... sitari, ... sitari'. This result represents the characteristic of the LH category whereby people tend to reply with several propositions, instead of the one specific solution, as is usually expected in the PC category. BA in the LH category includes 'anata' (you; 18) as a frequent expression, whereas NA in the LH category includes 'anta' (you; 14); the latter is a unpolite version of the former in Japanese. This result implies that answers with polite expressions are likely to be selected as a best answer. BA in PC includes the particle-adnominalizer 'no' (of; 2), particle-case 'wo' (-(objective), 3), and particle-case 'ni' (on; 4) as frequent expressions, while NA in PC includes the particle-adverbial 'ja' (-(conversational), 3), particle-adverbial 'nante' (-(conversational), 2), particle-final 'ne' (-(conversational), 7) and particle-final 'yo' (-(conversational), 14). The former results imply that BA in PC is a writing style like that found in instructionalmanualscontainingspecificexplanationswithclearandpertinenceexpressions('...wo

... ni ...'), while the latter results mean that NA in PC is a conversational style because all of the latter expressions are chatty ones. The results show that regarding the PC category, polite and 'manual-like' submissions are likely to be selected as the best answers than conversational submissions.

5. Conclusion

This study was an exploratory analysis of the stylistic characteristics of text submitted to Japanese Q&A communities. After observing the basic characteristics of the submissions, we applied principal component analysis and random forests to the text-feature matrix using the relative frequencies of function words. The results clearly show the stylistic characteristics of questions, best answers, and normal answers regarding the two categories of 'personal computers and peripheral devices' and 'love and human relations advice'. This study provided basic, but very important findings on how people differ in their communication styles regarding subject categories and on how people select communication styles online.

Our findings will be the foundation of various research. First, we will develop methods to distinguish the best answers from normal answers, by extending the findings of this study. Second, we will make a predictive transform or automatic paraphrase system focusing on the styles that we identified. Third, we will investigate the changing language usage of Japanese by comparing texts of the submissions against balanced corpora.

Acknowledgements

We were supported by Grant-in-Aid for Scientific Research 21800087 for Young Scientists (Start-up) from the Ministry of Education, Culture, Sports, Science and Technology, Japan, and a Mitsubishi Grant, 2008-2010. We used the Yahoo! Chiebukuro data provided by Yahoo Japan Corporation to the National Institute of Informatics. We would like to express our gratitude for the support. An earlier version of this study was presented at the 53th Annual Meeting of The Mathematical Linguistic Society of Japan at the Tokyo Woman's Christian University. We would like to thank the participants of that meeting for their helpful comments.

References

Aitchison J., and Lewis D.M. (editors) (2003). *New Media Language*. London: Routledge.

Argamon S., Whitelaw C., Chase P., Raj Hota S., Garg N. and and Levitan S. (2007). Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58 (6): 802-822.

Breiman L. (2001). Random forests. *Machine Learning*, 45: 5-23.

Garcia A.M. and Martin J.C. (2007). Function words in authorship attribution studies. *Literary and Linguistic Computing*, 22 (1): 49-66.

Grieve J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22 (3): 251-270.

Harper M.F., Raban D., Rafaeli S. and Konstan J.A. (2008). Predictors of answer quality in online Q&Asites. In *CHI '08: Proceedings of the Twenty-sixth Annual SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA. ACM, pp. 865-874.

Jin M. (2007). *R ni yoru Deta Saiensu*. Tokyo: Morikita Syuppan.

Jin M. and Murakami M. (2007). Authorship identification using random forests. In *Proceedings of the Institute of Statistical Mathematics*, 55 (2): 255-268.

Koppel M., Schler J. and Argamon S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60 (1): 9-26.

Miura A. and Kawaura Y. (2008). Why do people join Web-based knowledge-sharing communities? : Analysis on questioning and answering behavior. *Japanese Journal of Social Psychology*, 23 (3): 233-245.

Otsuka H., Inui T. and Okumura M. (2007). *Iken Bunseki Enjin: Keiryo Gengo-gaku to Syakai-gaku no Setten*. Tokyo: Colona Publishing.

Stamatatos E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60 (3): 538-556.

Suzuki T. (2009). Extracting speaker-specific functional expressions from political speeches using random forests in order to investigate speakers' political styles. *Journal of the American Society for Information Science and Technology*, 60 (8): 1596-1606.

Tuldava J. (1993). The statistical structure of a text and its readability. In Hfebicek, L. and Altmann, G., editors, *Quantitative Text Analysis*, Trier: Wissenschaftlicher Verlag, pp. 215-227.