Department of Social Systems and Management

Discussion Paper Series

No.1279

Facebook と Twitter の発言における特徴語の比較

（Comparing Specialized Vocabulary from Online Messages on Facebook and Twitter）

by

石井 健一
(Kenichi ISHII)

September 2011

UNIVERSITY OF TSUKUBA
Tsukuba, Ibaraki 305-8573
JAPAN

Facebook と Twitter の発言における特徴語の比較

石井健一（筑波大学大学院システム情報工学研究科）

要約　SNS(Social Network Service)として代表的な Facebook と Twitter から API の検索機能を用いて約 2 万件の発言を収集し、形態素解析により各 SNS の特徴語を抽出した。抽出された特徴語をみると、Twitter には顔文字や Twitter 特有の言い回しに関係した語が多かったのに対して、Facebook には広告やお知らせに関係する語が多く含まれていた。また、Twitter の発言に占める平仮名・カタカナの比率は、Facebook よりも有意に高かった。さらに、特定の検索語(家電製造の会社名と東京都の駅名)を用いて二つの SNS の特徴語を比較したところ、家電の会社名を含む発言では両者の特徴語は類似していたが、東京都の駅名を含む発言では Facebook と Twitter の特徴語はかなり異なっていた。

研究目的と背景

Facebook は、世界的には最も利用者の多い SNS であり、日本での登録者数は 2011 年 9 月時点で 472 万人[1]となっている。これに対して Twitter は対人関係をゆるくつなぐ「呟き」(最大文字数 140 文字)の SNS として使われている。日本では利用率や利用頻度の点で Twitter の方が Facebook よりも人気が高い（石井 2011b）。

本研究の目的は、Facebook と Twitter の発言内容から二つの SNS を比較することにある。現在、日本では多くの SNS が人気を博しているが、それらの利用実態は必ずしも明らかではない。本研究は、Facebook と Twitter における発言の特徴をみることによって利用実態の解明に役立つ資料を提供することを目ざす。また、発言データを実際に収集し比較することによって、クチコミ情報として SNS の発言データを用いることの可能性についても検討する。

本分析では、まず二つの SNS の全般的な発言内容(Facebook ではウォールへの書き込み)の傾向をみるため、「助詞」を使った検索結果を用いて Facebook と Twitter の発言内容を比較する。日本語の助詞はそれ自体意味のない語と考えられるので、助詞を用いた検索結果は、内容的なバイアスが小さいだろうと考えられる。次に個別のトピックで二つの SNS の差をみるため、「東京都内の駅名」および「家電ブランド」の検索語をもつ発言について Facebook と Twitter を比較する。これらの計量的的な分析のために、テキスト分析のツール kh-coder[2]を用いる。

方法

Facebook と Twitter の API(Application Program Interface)を利用して 2011 年 9 月上

---

[1] http://www.checkfacebook.com/

[2] kh-coder は樋口耕一氏によって開発されたソフトであり、ウェブ上で公開されている。なお、このソフトでは形態素解析に「茶筅」が使われている。http://khc.sourceforge.net/

旬に発言を自動的に収集した。Facebook には、Twitter のタイムラインに相当するようなリアルタイムで全発言をみる機能がない。そこで二つの SNS が API で提供しているキーワード検索機能を用いて発言データを収集した[3]。検索には、以下の 30 語を用いた。 (1)助詞(「は」「に」「を」「が」「で」「も」「と」)、(2)家電の主要会社名(「ソニー」「パナソニック」「シャープ」「東芝」)、(3)国名(「日本」「アメリカ」「中国」「韓国」「ドイツ」「イタリア」「フランス」「ロシア」「インド」)、(4) **東京都の駅名**(「池袋」「新宿」「渋谷」「上野」「六本木」「銀座」「秋葉原」「品川」「大手町」)。データの収集は、2011 年 9 月 4 日に行い、さらに当日、前日、前々日、三日前、四日前の五日間の発言をそれぞれ指定して各キーワードあたり 100 ずつ発言を収集した[4]。ただし、検索語によってはこの数の発言が得られなかったものもある。

日本語の助詞は単独では意味のない語なので、検索結果に与える内容上のバイアスは小さいと考えられる[5]。そこて、本論文ではまず、助詞の検索結果を用いて Facebook と Twitter の発言の全体的な比較を行う。その他の検索結果は、個別に結果をみることにする。

表 1　収集した発言数

	Facebook	Twitter
(1)　助詞	3302	2132
(2)　家電の会社名	320	2429
(3)　国名	3588	4272
(4)　東京都内の主要駅名	1036	4412
合計	8246	13245

(1)~(4)の各グループで収集した発言から重複した発言データを削除した[6]結果、収集した発言の個数は表 1 のようになった。Facebook において家電の会社名をキーワードとして含む発言は少なかった。なお、Facebook では各発言のタイトルが発言内容とは別のカテゴリーとして API から提供されるので、発言内容にタイトルを付加して発言データとした

[3] 用いた API は https://graph.facebook.com/search?q=”検索文字”&locale=ja_JP&type=post&limit=100 および http://search.twitter.com/search.json?lang=JA&q=”URL エンコードした検索文字"&result_type=recent&rpp=100&until=日付である。どちらも json 形式のデータが得られる。

[4] Twitter では過去 5 日、Facebook は 10 日程度までしか溯って検索できなかった。また、Twitter では各検索語あたり最大で 100 発言までしか表示しない。

[5] 他の検索方法として数字や記号を使うことも考えられるが、Facebook には日本人であっても英語で書いた発言があり、日本語の助詞を検索語に用いることでこうした発言を排除することもできると考えた。

[6] Facebook, Twitter ともにデータに固有の発言 ID があり、これで同一の発言かどうかを判別した。全く同一の文章でも発言 ID が異なれば、削除していない。

（Twitter にはタイトルはない）。

結果

(1) Facebook と Twitter の全般的な比較

表 2 は、助詞で検索して得た発言データについて Facebook と Twitter の記述統計量を比較したものである。一つの発言あたりの文字数は 140 文字という制限のある Twitter の方が短いのは当然であるが、意外なことに Facebook の方が一つの文が短い。一文あたり平均文字数は、Twitter では 39 文字[7]であり、Facebook では 23 文字である[8]。

表 2　記述統計による比較

	Facebook	Twitter
文の総数	33,816	26,200
発言数	8,219	13,157
発言あたり文字数平均	96.1	77.4
文あたり文字数	23.3	38.9
文字数（全角・半角）	789,444	1,017,944
ひらがな（全角）文字数	236,626	306,880
カタカナ（全角）文字数	78,457	126,375
半角（英数字・記号）文字数	176,300	261,149
半角カナの文字数	1,340	3,425
空白(半角・全角)文字数	27,098	32,329
一つの発言あたりのひらがな比率	34.2%	41.2%
一つの発言あたりのカタカナ比率	14.3%	18.9%

Twitter の発言には、平仮名と片仮名が多い。発言ごとに文字数に占める平仮名の比率を求めたところ、Facebook では平均 34.2%なのに対して Twitter では平均 41.2%[9]と 7%の差があり、両者には統計的な有意差が認められた(t=26.5; df=17133; p<.00001)。また、片仮名の比率についても同様の傾向がみられ、Facebook では 14.3%なのに対して Twitter では 18.9%であり、これも統計的な有意差が認められた(t=21.7; df=19887; p<.00001)。逆に言

[7] 全角、半角の文字をそれぞれ一文字として計算している（バイト数ではない）。この計算は perl で行なった。

[8] ただし、これは SNS の文章が通常の文章と異なるため、文の区切りを形態素解析のプログラムが適切に認識していない可能性がある。また、Facebook については、タイトルを発言データに加えたので、文が短くなる方向のバイアスとなっているであろう。

[9] この比率は、以下の式で計算した。ひらがな比率 = 発言の全角ひらがな数/(全文字数―空白数―半角文字数)　ただし、分母の文字数が 0 の発言を除く。

うと、Twitter では漢字を使う比率[10]が Facebook よりも低いと言える（後述する特徴語の分析の所でも同様の傾向がある）。

次に、各発言にリンク(URL)が含まれている比率を計算した（表 3）。Facebook の発言では 57.5%にリンクがある[11]のに対して Twitter ではこの比率は 36.5%であった。Facebook の方がリンクを含む発言が多いといえる[12]。また、発言の内容によってもリンクの比率は異なっていた。助詞で検索した得た発言データには、リンクが最も少なかった。これに対して家電関係の発言には、リンクが最も高い比率で含まれていた。これは、これらの発言は、外部の記事等を引用している比率が高いためと考えられる。

表 3　発言にリンクが含まれている比率(%)

	Facebook	Twitter
(1)　助詞	29.3	12.5
(2)　家電の会社名	84.4	68.3
(3)　国名	78.6	37.9
(4)　東京都の駅名	66.2	27.8
平均	57.5	36.3

次に Facebook と Twitter の発言内容の一般的な傾向を比較するため、助詞で検索して得た発言データを用いて、Facebook と Twitter の「特徴語」を求めた。特徴語とは、そのグループ（Facebook または Twitter）の発言において頻繁に使われるが、他のグループの発言ではあまり使われない単語[13]である。ここでは、Facebook と Twitter の出現頻度のズレを表 4 のようなクロス表における χ^2 値で計算し、χ^2 値が大きいほどその語の「特徴語」の傾向が強いとみなした[14]。

表 4　χ^2 の計算に使ったクロス表の例（χ^2=1091.0）

	Facebook	Twitter
「する」がある発言数	2909	999
「する」がない発言数	393	1303

10 ただし、本分析では全角の記号(※など)の個数を計算していないので漢字の正確な比率は出せない。

11 Facebook の標準機脳による画像のリンクは含まない。

12 中国語、英語の Facebook の発言と比較した結果では、日本の Facebook はリンクの比率が高かった（本論文では、分析結果は省略）。

13 分析の単位は、発言としている。Facebook の発言数 3302、Twitter の発言数 2132 のうち各単語が少なくとも一回以上出現した比率に対して $\chi 2$ 分析を行っている。

14 ただし、分析では各グループで出現頻度が 200 位以内に入る語のみに限定した。

表 5　Facebook と Twitter における全体的な「特徴語」(上位 28 個)

Facebook の特徴語				Twitter の特徴語			
語	Facebook 頻度	Twitter 頻度	χ^2 値	語	Facebook 頻度	Twitter 頻度	χ^2 値
する	2909	999	1091.0	RT	41	359	462.2
今日	1169	221	426.6	フォーゼ	4	49	63.6
なる	845	313	91.9	はる	30	80	52.8
−	216	26	86.2	ω	27	69	43.7
??	135	6	74.3	(>_<)	0	26	40.5
昨日	259	58	61.9	www	5	35	39.4
明日	137	16	54.7	ねる	42	80	36.3
写真	127	13	54.1	#	4	30	34.5
皆様	86	1	53.8	ゅ	4	29	33.0
年	116	10	53.0	ww	1	20	27.7
自分	233	54	53.0	あ	17	42	25.5
皆さん	108	9	49.9	(*	15	38	23.7
感じる	99	7	48.3	ライダー	4	22	22.6
感謝	88	4	47.8	W	31	55	22.4
おはよう	454	166	45.6	`)	13	34	21.8
アルバム	77	2	45.3	プリキュア	0	14	21.7
今朝	112	13	44.6	仝	28	49	19.5
日本	154	29	43.4	よね	0	12	18.6
フォト	73	2	42.7	わる	1	14	18.5
元気	119	17	41.8	∀	32	52	18.4
前	158	33	40.0	変身	0	11	17.1
頑張る	194	48	40.0	*)	7	22	16.4
できる	226	63	38.9	うん	3	16	16.2
本日	137	26	38.2	仮面ライダー	2	13	14.2
仕事	194	50	37.6	きる	10	24	14.1
月	172	41	37.1	QT	0	9	14.0
心	96	12	36.6	wwww	1	11	13.9
暑い	104	15	36.2	ww	7	20	13.8

表 5 をみると Facebook と Twitter の特徴語は、かなり異なることがわかる。Facebook の特徴語は、漢字の単語（今日、皆様、感謝、写真など）が多いのに対して、Twitter の特徴語には漢字の単語が少なく、平仮名（「はる」「ねる」「よね」など）や記号(「RT」[15]「WWW」「(>_<)」「Д」など)が多い。

Twitter で漢字の単語が少ないのは、Twitter で私的なやり取りが多いことを反映しているとみられる。また Twitter の特徴語に一般にはあまり使われない記号が多いのは、Twitter で顔文字が多く使われていて、これらの記号が顔文字を構成しているためである。表 6 にいくつかの顔文字の例を示したが、これらには∀やДといった文字が含まれている。また、Twitter の五番目の(>_<)は、(>_<)に相当する顔文字である(他の顔文字とは異なり、形態素解析でこの顔文字だけは一つの語として抽出された)。

一方、Facebook の特徴語には「皆様」「皆さん」「感謝」など Twitter に比べて「硬い」単語が多く含まれている。これは、Facebook が商業的・公的なコミュニケーションとして使われていることを反映しているとみられる。また、「する」「なる」という積極的な動詞がリストの上位にあるのも､個人や団体の活動のアピールが多いFacebookの発言の特徴を反映していると言えるかもしれない。

表 6　Twitter にみられた特殊な言い回しの例

・「ｗｗｗ」　(文末におき、鬱状態であることを示す)
・・・　今日やることまだあるのに徹夜フラグｗｗｗ
・顔文字
・・・　になれるよう頑張りたいです(*´∀｀*)(笑)
・・・　やっと見れる。･ﾟ･(*ﾉД`*)･ﾟ･。
・・・　やっぱり人多い(;´д｀)

(2) 家電の会社名を含む発言の特徴語

次に家電の会社名を含む発言[16]を Facebook と Twitter の発言全体(1-4 を全て含むデータ)の中かから取り出し、家電の会社名を含む発言の特徴語を抽出した。特徴語の選択基準としては$\chi 2$ 値[17]を用いた。$\chi 2$ 値は、各 SNS(Facebook または Twitter)での出現率と家電の

15　RT はリツイート(retweet)を示すもので、他の発言を転載していることを意味する。

16 「ソニー」「パナソニック」「シャープ」「東芝」に加えて、「Sony」「Panasonic」「日立」も家電ブランド名として抽出した。個別のブランド名の場合、出現数が少ないと結果が安定しないのでこのように「家電のブランド名」としてまとめた。

17 Dice 係数や相互情報量（中條・内山,2004）も検討したが、分析結果をみてχ^2値の結果が特徴語として最も適切と判断した。Dice 係数は、出現頻度の高い語が選ばれやすい傾向があり、相互情報量は出現頻度が少ないが一方のグループで出現しなかった語が選ばれやすい傾向があった。ただし、これらの指標については今後さらに検討する必要があろう。

会社名を含む発言での出現率での比率の差を示すものであり(ただし、表 7 では家電の会社名の出現率の方が高いもののみを選択している)、この差が大きいほど家電のの会社名を含む発言の特性をあらわす語であると見なすことができる。χ^2値は、Facebook と Twitter について別々に計算した。

特徴語の抽出結果(表 7)をみると、Facebook と Twitter の各 25 個の特徴語のうち 8 個が一致しており、両者は比較的似ているといえる。ただし、両者には違いもみられる。たとえば、Facebook のみの特徴語として「ニュース」「リリース」「速報」「公開」「独自」がある。これは、Facebook の発言に会社からの新商品紹介やニュースの引用が多いためと考えられる。

(3) 東京都の駅名を含む発言の特徴語

次に東京都の駅名（「池袋」「新宿」「渋谷」「上野」「六本木」「銀座」「秋葉原」「品川」「大手町」のいずれか）を含む発言を取り出し、(1)と同様に対応する特徴語を求めてみた(表 8)。この分析の目的は、身近な地名を含む発言にどういう特徴があるのかを Facebook と Twitter で比較することである。

特徴語の上位 25 語のリスト(表 8)をみると、25 個のキーワード中 4 個が一致しているが、家電の会社名の場合に比べて一致数が少なく、Facebook と Twitter の発言内容の差はより大きい。

- Facebook には、動詞が一つもない。これに対して、Twitter の特徴語には、「着く」「漏れる」という動詞が含まれている。
- Facebook の特徴語には、「出演」「OPEN」「場所」「徒歩」といった単語があり、イベント関係のお知らせが多いことがわかる。
- Facebook と Twitter の特徴語に「I」,「m」,「at」がある(ただし、Twitter には「m」はない)。これは、スマートフォンでのチェックイン通知機能（その場所にいることを自動的に知らせる）が使われているためであろう。たとえば「I'm at 恵比寿駅（Ebisu Sta.）（恵比寿南 1-5-5, 渋谷区）」は、恵比寿駅に来たというメッセージの例である。また、Twitter の「なう」は、「今・・にいます」の意味であるが、こちらは通常本人が書きこんだものである。
- 「佐賀」「玄海」「島根」「漏れる」という東京都の駅名と関係なさそうな単語が Twitter の特徴語にみられるのは、次のメッセージのリツイート(転送)がデータの中に多数(44 回)含まれていたためである。「福井県の原発銀座，島根県の島根原発，佐賀県の玄海原発では，いずれも少し高い値を示していることがわかる。事故がなくても漏れている」。ごく一部の利用者の発言が繰り返してリツイートされることによって、特徴語の結果に影響を与えていることを示している。

表 7　家電の会社名を含む発言の特徴語

Facebook					Twitter				
	特徴語	全体での頻度	家電会社の発言での頻度	χ^2		特徴語	全体での頻度	家電会社の発言での頻度	χ^2
1	液晶	38	34	667.7	1	液晶	165	160	334.2
2	Android	39	30	537.5	2	タブレット	152	147	305.7
3	ウォーク	29	25	479.3	3	発売	212	152	232.3
4	ベース	33	25	443.1	4	搭載	134	118	223.5
5	タブレット	41	27	437.2	5	3D	97	94	194.2
6	対応	76	33	394.3	6	テレビ	208	132	172.8
7	発表	113	36	331.1	7	価格	127	96	154.3
8	統合	26	19	328.6	8	統合	74	73	152.8
9	速報	55	25	308.8	9	Android	93	81	150.8
10	α	18	15	281.1	10	事業	95	78	136.7
11	公開	78	27	266.1	11	ディスプレイ	67	64	129.9
12	発売	100	30	260.4	12	発表	241	127	127.5
13	アプリ	36	18	238.7	13	アクオス	56	56	118.3
14	ファインダー	11	11	228.6	14	端末	80	66	116.0
15	EL	14	12	228.5	15	中小	56	54	110.4
16	リリース	28	15	209.2	16	iPad	83	64	104.5
17	独自	15	11	190.2	17	パネル	53	51	104.0
18	有機	19	12	188.0	18	対応	121	78	103.1
19	画素	9	9	186.9	19	製品	97	65	90.0
20	マーケットアプリ	9	9	186.9	20	国内	114	71	89.6
21	中小	18	11	168.4	21	無線	77	57	88.8
22	2430	8	8	166.1	22	カード	92	62	86.4
23	異彩	8	8	166.1	23	三菱	43	41	82.8
24	形状	12	9	157.8	24	デジカメ	40	39	80.3
25	NEX-7	9	8	155.5	25	SDHC	37	37	78.0

表 8　東京都の駅名を含む発言の特徴語

Facebook					Twitter				
	特徴語	全体での頻度	東京都の駅名での頻度	χ^2		特徴語	全体での頻度	東京都の駅名での頻度	χ^2
1	At	127	97	203.8	1	東京	421	319	51.8
2	東京	262	128	158.5	2	ヒルズ	98	98	32.8
3	m	126	84	152.2	3	ライブハウス	91	90	29.3
4	)	837	256	142.8	4	出勤	107	96	24.9
5	I	149	87	134.9	5	10:00	75	74	23.9
6	@	86	62	121.5	6	At	87	81	22.9
7	(	813	228	100.5	7	店	137	112	22.5
8	ライブ	93	58	96.5	8	ヘル	82	76	21.3
9	w/	36	34	85.6	9	I	89	80	20.8
10	(@	45	37	82.2	10	イヤ	69	65	18.9
11	Live	62	43	80.4	11	本日	233	162	18.3
12	賃貸	36	31	71.8	12	E	73	65	16.4
13	出演	68	42	68.7	13	ホン	64	59	16.3
14	土	90	48	65.3	14	佐賀	49	49	16.3
15	Sta.)	28	26	64.4	15	漏れる	48	48	15.9
16	ビル	50	34	62.0	16	なう	277	182	15.8
17	OPEN	37	29	61.3	17	オープン	89	74	15.5
18	/	195	73	56.7	18	福井	46	46	15.3
19	others	25	23	56.4	19	未遂	46	46	15.3
20	場所	100	47	53.2	20	着く	105	83	14.9
21	START	27	23	52.6	21	島根	48	47	14.9
22	Bar	29	22	44.9	22	(@	79	67	14.8
23	徒歩	32	23	44.4	23	玄海	44	44	14.6
24	当日	53	30	43.9	24	注 1	43	43	14.3
25	in	130	51	42.8	25	TSUTAYA	49	47	14.2

注 1 Twitter の特定 ID だったので表から削除した

結論

特徴語の全体的な比較から、Facebook に比べて Twitter では私的なやりとりの発言が多いという傾向が確認できた。これに対して、Facebook は公的なお知らせや企業広告とみられる内容が相対的に多い。このことは Twitter の方がクチコミ情報源として相対的に優れたデータソースであることを示しているようであるが、Twitter にも販売促進を目的としたお知らせやサクラとみられる発言が含まれていないわけではない。特定の商品の販売促進を目的とした発言や Bot と言われるコンピュータによる自動的な書き込み（石井、2011a）の影響をどのように排除するかが、今後クチコミのデータとして分析する場合の課題となるであろう。

形態素解析は、大量のテキスト分析を自動化するための有効な手法ではあるが、現時点ではいくつかの方法論的な問題点がある。その一つは、顔文字や Twitter の特殊な言い回しなど、既存の形態素解析のプラグラムでは、十分に対応できていない表現が多くあるということである。SNS の発言をクチコミのデータとして本格的に分析するためには、こうした特有の言い回しを考慮する必要があろう。また、もう一つの問題点は、個別の会社名やブランド名は、SNS にはあまり多く出現しないことである。今回の分析でも、個別の家電の会社名がある Facebook の発言は少なかった。また、東京都の駅名の分析では、複数の駅名の検索結果を合併したにもかかわらず、多数回リツイートされた発言の影響が見られた。つまり、特定の検索語に依存した分析は、少数の発言者の発言によって強く影響を受けてしまう可能性があると言えよう。

参考文献

石井健一（2011a）. マイクロブログ Twitter における日本人利用者の特徴、Department of Social Systems and Management Discussion Paper Series, No.1277, pp.1-7. http://www.sk.tsukuba.ac.jp/SSM/libraries/pdf1276/1277.pdf

石井健一（2011b）.「強いつながり」と「弱いつながり」の SNS―利用と満足の視点からみた 5 つのソーシャル・ ネットワーキング・サービスの比較―、情報通信学会大会個人研究発表配布資料 2011 年 7 月 3 日(専修大学)、http://www.jotsugakkai.or.jp/doc/taikai2011/K1Ishii.pdf

Ishii, Kenichi (2008). Uses and Gratifications of Online Communities in Japan. *Observatorio*, 2(3), pp.25-37.

Ishii, Kenichi & Ogasahara, Morihiro (2007). Links between Real and Virtual Networks: A Comparative Study of Online Communities in Japan and Korea. *CyberPsychology & Behavior*, 10(2), pp.252-257.

中條清美・内山将夫（2004）「統計的指標を利用した特徴語抽出に関する研究」『KATE Bulletin』18, pp.99-108. http://www5d.biglobe.ne.jp/~chujo/data/KATE2003.pdf