自 然 言 語 処 理　130－7
（1999．3．4）

文末態度表現に注目したWeb Pageの調査

土井 晃一*

学術情報センター
〒112-8640 東京都文京区大塚 3-29-1
(03)3942-7967
doy@rd.nacsis.ac.jp

和文抄録

近年、インターネットによる検索が盛んに行われるようになった。しかし、検索エンジンによる検索は、しばしば適切な結果をもたらさない。一般の検索エンジンでは、いわゆるキーワードを主体とした検索が主流である。普通、付属語は検索では利用されない。本論文では、普通、検索には使われない付属語、その中でも特に文末表現に話を絞る。文末表現は文章のジャンルと関係が深いので、文末表現を研究することで、文章のジャンルがわかる可能性がある。本論文では、文末表現を網羅し、どんなジャンルのweb pageが抽出できそうかを調べ、さらに、web page上での文末表現についての基礎統計を調べる。検索の結果、日記などのジャンルのweb pageが抽出できそうな見とおしを得た。

On the Categorization of Web Pages based on Attitude Expressions at the ends Sentences.

Kouichi DOI[1]
National Center for Science Information Systems (NACSIS)
3-29-1 Otsuka, Bunkyo-ku Tokyo 112-8640 JAPAN
+81-3-3942-7967

Abstract

Nowadays, one can search the internet very easily. The results, however, often are very poor. Most search engines retrieve data using so called “key words”. Stop words are rarely used by search engines. In this paper, we use stop words, especially, the expressions at the end sentences. To characterize web pages the expression at the end of a sentence and the genre of the sentence are closely related. It is possible to discern the genre of a sentence from the expression at its end. In this paper, we tally the expressions at the ends of sentences exhaustively. From a statistical analysis of the expressions we estimate the genre of the web pages. In paticular, we estimated the likelihood of a web page being a diary.

*(株) 富士通研究所からの客員助教授
[1]transferred from Fujitsu Laboratories Limited

1 はじめに

近年、インターネットによる検索が盛んに行われるようになった。しかし、検索エンジンによる検索は、しばしば適切な結果をもたらさない。一般の検索エンジンは、

1. いわゆるキーワードを主体とした検索
2. ディレクトリ・サービス

などを主体としている。普通、付属語 (助詞・助動詞など) は不要語として無視しているようである。本論文では、普通検索には使われない付属語、その中でも特に文末表現に話を絞る。

文末表現は文体と関連する。文体は文章のジャンルと関連する。つまり、文末表現を研究することで文章のジャンルがわかる可能性がある。

本論文では、

1. 文末表現を網羅すること
2. どんなジャンルの web page が抽出できそうか調べる。
3. web page 上での文末表現についての基礎統計を調べる。

ことを目的にする。

以下、第二節では、検索環境について述べ、web page での句点の使われ方について調べる。第三節では、文末表現の洗い出しを行い、その文末表現でどのようなジャンルの web page が抽出できそうかを調べる。第四節では、ムードによる検索を行い、基礎統計量を取る。最後に、第五節では、全体のまとめをする。

2 検索環境

以下の検索は Livelink search 検索エンジンを用いて行なった。検索対象は ne ドメインの web page とイメージデータからなる　web page を除くほぼすべての日本の web page である。対象となる web page の数は 608,983 ページである。

まず、句点の使い方を調べた。全角の「。」が使われているページが 423,440 ページ、半角の「。」が使われているページが 5,230 ページ、全角の「．」が使われているページが 51,144 ページ、半角の「.」が使われているページはなかった。以下、検索の都合上、検索対象を全角の「。」が使われているページに話しを絞る。

3 文末表現の洗いだし

日本語の文末表現を洗い出し、その文末表現がどのような web page に用いられているかを調べた。

検索はあ。、い。...のように「あいうえお」順で行い、さらに、が。ぱ。などの濁音・半濁音・促音についても行った。

文末表現で一番種類が多いのは、体言止であると思われる。しかし、体言止は種類が限りなく存在し、それらを統一して扱うことが困難であるので、基本的に対象からはずした。また、明らかな書き間違いを対象からはずした。その結果、文末表現として抽出されたのは、助詞、助動詞、典型的な用言、感動詞が主になった。

ここで「典型的な用言」とは、きわめて出現頻度が高いか、あるいは、ある種のジャンルの文に特徴的に現れる用言を指す。

3.1 あ。とお。

あ。で終わる文末表現には、

1. 元々の文末表現があ行で終わるため、その「あ」を表記したもの
2. 感動詞的に使われているもの

の主として 2 種類のものが見受けられた。1. は話し言葉をそのまま表記したもの、あるいは、文体を口語体にしたいために用いられていると考えられる。1.2. 共に話し言葉に特徴的な現象であるためか、日記・対談・独白・くだけた座談会などによく見受けられた。

抽出された文末表現は、それぞれ、

1. なあ。だあ。そっかぁ。ないですかぁ。なぁ。これじゃあ。だったあ。ですかあ。すうよりもさあ。ましたあ。とってきてやあ。満足ぢゃあ。生きなくっちゃあ。なーあ。回られちゃあ。あのなぁあ。古いんだからあ。楽しいなぁあ。がんばらねばあ。
2. あ。ひゃあ。まあまあ。ああ。じゃあ。まあ。あーあ。わあ。はあ。きゃあ。あちゃあ。はあ。ははあ。あらあ。あ〜あ。あっちゃあ。があ。やあ。さあ。はあはあ。よっしゃあ。ぐはあ。

あ～～あ。さあさあ。おっしゃあ。そりゃあ。うあ。

である。1. のもとの言語表現は、

な。だ。か。では。た。さ。や。わ。じゃ。ちゃ。だから。ば。

であることが容易にわかる。また、1.2. の両方に含まれる文末表現としては、

うわぁ。あらぁ。いやぁ。まさかぁ。いやあ。わあ。うわあ。わはははあ。であであ。あらあ。

が抽出された。新しく抽出された文末表現としては、

にゃあ。ニャあ。行ってみんしゃあ。あらざあ。知ってらあ。お仕事ガンバんなきゃあ。

がある。

この傾向はお。にも見受けられた。それぞれ、

1. やられたよお。みましょうお。お願いしますよぉ。なるほどお。でもぉ。きっとお。
2. お。おお。おおお。

が見受けられた。

3.2 あ。とお。以外

あ。とお。以外で終わる文末表現について、表1と表2にまとめておく。表中、第一列目は文末表現の見出しを示す。この見出しを種に実際の検索を行なった。第二列目は抽出された文末表現である。第三列目は多数見受けられた文法要素である。しばしばこの文法要素によって、検索を打ち切りざるを得なかった。この文法要素だけを適切に排除する方法が存在しないからである。第四列目はその文末表現が典型的に現れる web page のジャンルを示した。これは全体をざっと見渡した結果、著者が直観的に多いと感じた web page のジャンルである。

この検索から、抽出できそうな web page のジャンルを挙げると、

機械の販売・数学の問題・独白・古文・関西弁・法律・方言の感じを出したもの・報告・きわめてまじめなもの・きわめて口語的なもの・Q&A・依頼・日記・行政関係・議事録・マニュアル・申込書・提言書・座談会・株式関係・辞書・項目説明・発話録・e-mail・対談・物語・link 先・窓口の案内・スローガン

が挙げられる。このうち重要だと思われるのは、

独白・極めて口語的なもの・日記・e-mail

である。これらは個人の主観的な情報が数多く盛り込まれている。例えば、何かの商品のことを知りたいと思ったときに、その商品の販売もとの home page を見れば、スペックなどの公式の情報は得られるが、その商品の便利さ、有用性など使った人にしたわからない情報は得がたい。日記などの web page には後者の情報があふれているので役立つということがある。日記などの web page を検索できるようになることは、このような点からも重要と考えられる。

また、逆に、百科事典的な情報を得たいのに、日記のようなものばかり出てきて困るというときフィルターとして利用することが考えられる。

抽出できそうな web page のジャンルということは、精度が高いことが予測される。本当に精度高く抽出するにはどうしたらよいか、さらに、再現率はどうなっているのかが今後の課題である。

4 ムードによる検索

まず、日本語の文法書 [1] のムードの章の例文から、文末表現を抽出した。その各々の文末表現について、検索サーバを動かし、

1. その文末表現を含む web page の数
2. その文末表現の総数
3. その文末表現を含む web page での「。」の数

を検索した。

表3と表4に結果を示した。表中、第一列目はムード、第二列目はそのムードの取る文末表現を挙げた。第三列目は対応する第二列の言語表現が現れる web page の数、第四列目はその言語表現の総数を示す。第五列目は、対応する第二列の言語表現が現れる web page での「。」の数を示す。第六列目は第四列目を第三列目で割った値、第七列目は第五列目を第三列目で

文末表現の見出し	文末表現	備考	web page　のジャンル
い。	しれない。気づいていない。ちがいはない。思ってはいない。みずみずしい。ください。下さい。やすい。いい。よい。		
	しなさい。みなさい。	命令形	マニュアル・申込書
う。		動詞の終止形	
え。	ええ。ねえ。救え。揃え。やっちまえ。		
	従え。使え。思え。	命令形	マニュアル
お。	逃げちゃお。にゃお。もお。	前述のものは除く	
か。	ではないか。どうか。いるか。	疑問	提言書
	どこですか。		提言書
き。	べき。すき。好き。		
	するとき。		法律
	マニュアル付き。		機械の販売
	高校生以上向き。		数学の問題
く。		動詞の終止形	日記
け。		動詞の命令形	
	意味がないわけ。		
こ。	ここ。とこ。あそこ。		
さ。		終助詞のさ。	座談会
し。	なし。久し。べし。如し。多し。		
	気分でもあるし。あったし。		座談会が多い・口語体の文章
す。	です。ます。示す。表す。		
せ。	示せ。	動詞の命令形	
	ご覧じませ。		古文
そ。	こちらこそ。そそ。(あいづちの「そうそう」の省略形)		
た。		過去のた	
ち。	頭打ち。		株式関係
	人たち。		
つ。	の一つ。持つ。		
て。	思って。思ってて。	動詞のテ形	
	はじめまして。		e-mail
と。	すること。したこと。		法律
	のこと。		辞書・項目説明
な。	だな。かな。よな。ですな。		きわめて口語的
に。	ありますように。目標に。だろうに。		
ぬ。	できぬ。		古文
ね。		終助詞のね	口語的な日記・座談会
の。	混じっているの。	形式名詞	口語的なもの
	以下のもの。		硬い文章
は。	というのは。ははは。こんにちは。		発話録・e-mail
ひ。	我思ひ。		古文
	ぜひ。		対談
ふ。	見たまふ。		古文
	ふふ。	笑い	対談・e-mail

表 1: 抽出された文末表現 (その 1)

へ。	いわさきさんへ。		e-mail
ほ。	とほほ。ほほほ。		日記
ま。	ゆりこさま。		e-mail・古文
み。	休み。のみ。見込み。		情報・日記
む。	を含む。		法律
		動詞の終止形	日記
め。	お勧め。		情報
も。	等の話題も。		情報
	ここも。		リンク先
	けれども。けども。		発話録
	よかったのかも。		日記
や。	思うんや。	関西弁	
ゆ。	見ゆ。		古文
よ。	猫よ。やろうよ。		物語・発話録
ら。	かしら。から。ながら。		
り。	至れるあり。けり。なり。		古文
	あり。有り。より。		
る。	である。している。する。できる。		
	戻る。		link 先
れ。	売り切れ。生まれ。		
		動詞の命令形	
ろ。		動詞の命令形	
	だろ。ところ。7時ごろ。覚えたてのころ。		
わ。	ですわ。ますわ。	終助詞のわ	
	こんばんわ。こんちわ。でわ。		
を。	問い合わせを。	格助詞のを	窓口の案内・スローガン
ん。	ません。片岡さん。くん。だよん。まんねん。		
が。	ですが。ますが。	接続助詞のが	
ぎ。	すごすぎ。かっこよすぎ。		
ぐ。	下車すぐ。		案内
		動詞の終止形	
げ。	打ち上げ。引き下げ。格下げ。		
ご。	りんご。いちご。		
ざ。	わざわざ。		
じ。	同じ。		法律
		禁止のじ	古文
	感じ。		
ず。	あらず。		古文
	変わらず。はず。あしからず。		
ぜ。		終助詞のぜ	
ぞ。		終助詞のぞ	
だ。		判定詞のだ	
づ。	多いはづ。	わざとあるいは無意識に間違える	
で。		接続助詞ので	発話録
ど。	など。けど。		
		終助詞のど	
ば。	さらば。ならば。		
び。	並び。高い伸び。		
ぶ。	呼ぶ。学ぶ。		
べ。	調べ。選べ。		
ぼ。	めいぼ。		
ぱ。	やっぱ。		
ぺ。	やっぺ。	田舎の感じを出す	
っ。	とさっ。	きわめて口語的	

表 2: 抽出された文末表現 (その 2)

割った値、第八列目は第四列目を第五列目で割った値をそれぞれ示す。つまり、第六列目はある言語表現が現れる web page の中で何回その言語表現が現れるかを示すことになる。第七列目はある言語表現が現れる web page の平均文数、第八列目はある言語表現が現れる web page で、その言語表現が占める文の割合をそれぞれ示すことになる。また、件数とはヒットした web page の数、ワード数とはヒット総数を表す。

これらの表を見てわかることをまとめておく。まず、第六列目を見る。ここで数値の大きいものを大きい順に挙げると、

> 意思の「ます。」(5.69)・禁止の「ない。」(3.81)・確言の「する。」(3.38)・疑問の「か。」(3.22)

が挙げられる (括弧の中の数値には実際の値を挙げた。以下同様。)。これらの言語表現は一度使われると、何度も使われる傾向が高いことがわかる。また、逆に数値の小さいものは件数・ワード数共に極めて小さいものが多いので、ここでは考察の対象外とする。

次に、第七列目を見る。ここで数値の大きいものを大きい順に挙げると、

> 説明の「わけですか。」(417.86)・禁止の「だめだ。」(330.72)・勧誘の「てみないか。」(310.50)・当為の「んじゃなかった。」(309.76)

が挙げられる。いずれも件数は少なめだが、これらの言語表現が現れると文の数が極めて大きくなることがわかる。逆に数値の小さいものは、数値の小さい順に挙げると、

> 意思の「ます。」(21.91)・説明の「のです。」(57.15)・確言の「する。」(64.98)・申し出の「ましょう。」(69.43)

が挙げられる。これらの言語表現が現れると文の数が極めて小さくなることがわかる。

最後に、第八列目を見る。ここで数値の大きいものを大きい順に挙げると、

> 意思の「ます。」(25.98%)・確言の「する。」(5.20%)・禁止の「ない。」(4.96%)・説明の「のです。」(4.18%)

が挙げられる。ひとたびこれらの言語表現が現れると、何度もその言語表現が現れることがわかる。逆に、数値の小さいものを小さい順に挙げると、

> 許可の「いいかい。」(0.19%)・説明の「わけですか。」(0.26%)・勧誘の「てみないか。」(0.32%)・当為の「んじゃなかった。」(0.36%)

が挙げられる。これらの言語表現は、使われている件数・ワード数共に少なく、また、使われても非常にまれであることがわかる。

さらに分析すると、

> 確言の「する。」・禁止の「ない。」・意志の「ます。」

は一度使われると何度も使われ、しかも、文の数が少なくなりがちである。

今後の課題は、

1. さらに数値的な分析をする。
2. これらの文末表現が実際にどのように使われているかを調べる。
3. ある web page のジャンルでどのように文末表現が使われるかをこのデータと比較する。

である。

5 おわりに

本論文では、文末表現に注目して web page の調査を行った。まず、文末表現を洗い出し、その文末表現がどのような web page に用いられているかを調べた。次に、ムードに着目して、ムードを示す文末表現についてその使われ方について調べた。

今後は、さらに web page のジャンルと文末表現の関係について調べていきたい。

謝辞

有益な助言を頂いた、富士通研究所 渡部勇様と学術情報センターの相澤彰子先生に深謝致します。

参考文献

[1] 益岡隆志, 田窪行則. 基礎日本語文法 - 改定版 -. くろしお出版, 1995.

		A	B	C	B/A	C/A	B/C
		件数	ワード数	ワード数			
確言	する。	41599	140643	2703143	3.380922618	64.98096108	0.052029434
	ではない。	8285	14110	1012123	1.703077852	122.1633072	0.013940993
	だね。	1588	2915	283321	1.835642317	178.413728	0.010288683
命令	よ。	16978	47146	1720828	2.776887737	101.3563435	0.027397276
	て。	11133	22642	1400527	2.033773466	125.7996048	0.016166772
	てよ。	333	400	78806	1.201201201	236.6546547	0.005075756
	こと。	12839	38078	1398336	2.965807306	108.9131552	0.027230937
	んだ。	7775	15179	980556	1.952282958	126.1165273	0.015479993
	のだ。	10885	27475	1075277	2.524115756	98.785209	0.025551556
禁止	な。	9340	21994	1241348	2.354817987	132.9066381	0.017717836
	ない。	40459	154047	3103133	3.80748412	76.69821301	0.04964241
	なよ。	243	300	68338	1.234567901	281.2263374	0.004389944
	いけない。	1713	2389	338044	1.394629305	197.3403386	0.007067127
	だめだ。	118	141	39025	1.194915254	330.720339	0.003613069
許可	いいよ。	516	1274	116428	2.468992248	225.6356589	0.010942385
	かまわない。	121	131	24771	1.082644628	204.7190083	0.005288442
	いいかい。	8	9	4844	1.125	605.5	0.001857969
依頼	ちょうだい。	168	184	29472	1.095238095	175.4285714	0.006243214
	もらえますか。	19	31	2622	1.631578947	138	0.011823036
	いただけないでしょうか。	95	115	29136	1.210526316	306.6947368	0.003947007
	もらいたい。	771	1063	137111	1.378728923	177.8352789	0.007752843
	んですが。	763	1070	199610	1.402359109	261.6120577	0.005360453
当為	べきだ。	735	1109	128921	1.508843537	175.4027211	0.008602167
	といけない。	318	385	70357	1.210691824	221.2484277	0.005472092
	べきだった。	100	115	28110	1.15	281.1	0.004091071
	ものだ。	3386	4650	411148	1.373301831	121.4258712	0.011309796
	ほうがいい。	179	200	36659	1.117318436	204.7988827	0.005455686
	方がいいぜ。	2	2	300	1	150	0.006666667
	ことだ。	3361	4909	406203	1.460577209	120.8577804	0.01208509
	んだった。	252	281	58950	1.115079365	233.9285714	0.004766751
	のだった。	2086	3304	299847	1.583892617	143.7425695	0.011018953
	んじゃなかった。	17	19	5266	1.117647059	309.7647059	0.003608052
	のじゃなかった。	3	3	264	1	88	0.011363636
意志	ます。	276839	1575746	6065778	5.69192202	21.91085071	0.259776405
	つもりです。	1839	2092	267919	1.137574769	145.6873301	0.00780833
	よう。	8621	13995	1053667	1.623361559	122.220972	0.013282185
相手の意思を尋ねる	ますか。	1908	4812	301645	2.522012579	158.0948637	0.015952527
	つもりですか。	9	9	833	1	92.55555556	0.010804322
申し出	ましょう。	27342	43896	1898445	1.605442177	69.43328944	0.023122081
	ましょうか。	477	643	122160	1.348008386	256.1006289	0.005263589

表 3: 文末表現 (その 1)

勧誘	ませんか。	4200	5292	483312	1.26	115.0742857	0.010949449
	てみないか。	10	10	3105	1	310.5	0.003220612
提案	どうだ。	74	79	14489	1.067567568	195.7972973	0.005452412
	どうか。	1072	1778	215540	1.65858209	201.0634328	0.008249049
願望	たい。	12938	25975	1474413	2.007651878	113.9598856	0.017617181
	てほしい。	2113	3724	299145	1.762423095	141.573592	0.012448812
	てほしいな。	51	55	14550	1.078431373	285.2941176	0.003780069
概言	だろう。	11616	24864	1253565	2.140495868	107.9170971	0.019834632
	のでしょう。	2646	3785	440207	1.430461073	166.366969	0.008598228
	でしょう。	16784	34800	1601155	2.073403241	95.39770019	0.021734311
	まい。	1631	2086	299813	1.278969957	183.8215819	0.00695767
	らしい。	6762	12663	887529	1.872670807	131.2524401	0.014267703
	ようだ。	5134	7962	587841	1.550837554	114.4996104	0.013544479
	らしかった。	207	234	52227	1.130434783	252.3043478	0.004480441
	ようでした。	680	827	140978	1.216176471	207.3205882	0.005866164
	はずです。	3275	4315	445337	1.317557252	135.9807634	0.009689291
	はずですよ。	27	30	5146	1.111111111	190.5925926	0.005829771
	はずだ。	1531	1924	215961	1.256694971	141.0587851	0.008909016
	はずなんだが。	4	4	1028	1	257	0.003891051
	かもしれない。	5625	8747	780194	1.555022222	138.7011556	0.011211314
	にちがいない。	215	248	28713	1.153488372	133.5488372	0.008637203
	そうだ。	4617	6987	546018	1.513320338	118.2625081	0.012796281
	そうです。	9071	17585	1051410	1.938595524	115.9089406	0.01672516
	そうだった。	417	459	78668	1.100719424	188.6522782	0.005834647
	そうではない。	239	261	66429	1.092050209	277.9456067	0.003929007
	という。	11745	23570	879710	2.006811409	74.90080885	0.02679292
	とのことだ。	262	295	21255	1.125954198	81.1259542	0.013879087
	ということです。	3177	5152	447891	1.62165565	140.9792257	0.011502799
	ということでした。	476	532	95896	1.117647059	201.4621849	0.005547677
	とのことだった。	99	110	15500	1.111111111	156.5656566	0.007096774
説明	んだ。	7775	15179	980556	1.952282958	126.1165273	0.015479993
	のだ。	10885	27475	1075277	2.524115756	98.785209	0.025551556
	のです。	48219	115267	2755674	2.390489226	57.14913208	0.041828968
	んです。	8608	19238	1045520	2.23489777	121.4591078	0.018400413
	んですね。	15403	18113	1166621	1.175939752	75.73985587	0.015526036
	のですね。	1588	2309	321529	1.454030227	202.4741814	0.007181312
	わけです。	3713	8697	490440	2.3423108	132.087261	0.017733056
	わけですか。	42	45	17550	1.071428571	417.8571429	0.002564103
	わけだ。	1596	2343	215316	1.468045113	134.9097744	0.010881681
	わけではない。	1836	2272	292470	1.237472767	159.2973856	0.007768318
比況	のようだ。	980	1152	147191	1.175510204	150.194898	0.007826565
	ようだ。	5134	7962	587841	1.550837554	114.4996104	0.013544479
疑問	か。	33546	107961	2878316	3.218297263	85.80206284	0.03750839
	のだろう。	2804	3911	422373	1.394793153	150.632311	0.009259588
	のかな。	1077	1476	249570	1.370473538	231.7270195	0.005914172
	かね。	1316	1980	302366	1.504559271	229.7613982	0.006548355
	んですか。	614	1062	176247	1.729641694	287.0472313	0.006025634
	のですか。	687	1276	138183	1.857350801	201.139738	0.009234132
	の。	10908	19003	1180904	1.742115878	108.2603594	0.016091909
平均					1.64343	171.3881	0.014982
メジアン					1.394793153	141.573592	0.009234132

表 4: 文末表現 (その 2)