機能語句の話し言葉らしさ指標

玉城 伸仁　　黒橋 禎夫
京都大学大学院情報学研究科
{tamaki, kuro}@nlp.kuee.kyoto-u.ac.jp

1　はじめに

話しことばを計算機で扱おうと考えたとき、その書き言葉とは違ったいくつかの特徴に注意を払わなければならない。ノンバーバルなコミュニケーションとの連動はもちろんのこと、言語的な内容においても、より平易な語彙、単純な構文が好んで用いられる。言い間違いや、言いよどみ、言い直しが頻繁に発生する。しばしば文法的な適確性を欠いていたり、一文としては中途半端な言い差しの文が意図的に使用されたりするなど種々の特徴をもつ。これらの特徴は話しことばが基本的に音声媒体によって伝達されるものであることに起因する特徴と、典型的には対人的な相互行為である会話の場で使用されるものであるということから来る特徴に整理できると考えられる。対話応答などを念頭においた計算機による言語生成システムにおいては、後者の対人的という性質が中心的な課題であると考えられる。情報技術の社会的な応用を考えるとき計算機は人に「語りかけ」ねばならない。

電子メディア上でしばしば観察されるくだけた文体はとても話しことば的であると感じられる。一方で、学術的な講演などで用いられる口調からはむしろ書き言葉的であるという印象を受ける。このような感覚は談話分析の分野で involvement という概念で説明されてきた (Besnier, 1994 [1];Maynard, 2000 [4])。会話的であるか文章的であるかの区別は、会話が相手と共同で作り上げる「行為」であって、常に相互作用を予定した「場」への働きかけとして実現されるものであるということから生じる。ここでは involvement の用語を場への働きかけをともなう言語表現上のストラテジーの意味に限定して用いる。私たちはこういった自然言語によるコミュニケーションの仕組みを計算機が扱える明確さで記述していきたいと考えている。

場への働きかけを行う代表的な方法は情意 (affection) を表出することである。Cook(2001) [2] は affect keys の例として日本語の終助詞などをあげている。情意表出の職能は程度の差はあれ多くの語句に備わっているものと思われるが、談話を制御するストラテジーとしては文の内容的成分よりも機能的成分の寄与が重要であろうと考えられる。これは内容語の多くが文の命題構成に関わり、選択の自由が強く制限されるのに対して、機能語句は文の意味内容とはある程度独立に制御可能なためである。こうした特徴が機能語の選択に書き手の好みを反映させ、いわゆる口調・文体を形成していると考えられる。金 (2002) [3] は機能語句として助詞の選択に注目し、その選好パターンにより高精度に著者判別が可能であることを報告している。

ところで機能語句に注目した多変量解析は書き言葉と話しことばの判別でも有効に働くと考えられるが、文生成や言い換えといった応用を念頭においた場合少々扱いにくい。判定する単位が文章単位であるということと、判別過程の透明性が低く、不適切な出力であると判別できた後に、どのような変更を加えれば適切になるのかという情報が得にくいということが問題である。

そこで本稿では、ある程度の精度で一文ごとの口調判定が可能であって、基準が透明で言語的な意味が理解しやすい話しことばらしさの尺度を提案する。

はじめに、会話文を構成する機能語句群は密接に連携しあって働いているものであり、会話文とみなせる文において共起しやすいと予想した。出現分布の偏りを調べるためには異なる条件下での出現率の違いを調べれば良い。今回は、終助詞を含む文は多くが会話文であるということを手がかりとした。機能的形態素の WWW コーパスにおける出現率と会話文に限定したときの出現率の比を計算した。これを各形態素の口語性を反映した指標であると考える。次に、この指標を使って各文に得点を与え、会話文らしさを評価した。形態素の口語性をもって文の会話文性を評価し、文の会話文性をもって形態素の口語性を逆評価する。これを交互に繰り返す一種の半教師あり学習を行った。得られた指標はコーパスの特性に依存し、恣意的な開始条件の影響は相対的に小さいものであると考えている。

2　用いた資料と指標の計算方法

2.1　コーパス

WWW から収集し形態素解析器 JUMAN で解析した 5 億文の中から、調査対象となる形態素を 3 つ以上含む文を抽出した。文数は 3.7 億文であった。

2.2　調査対象とした形態素

JUMAN の辞書で以下のように分類される形態素と動詞・形容詞の活用語尾を調査対象とした。計数する際にレンマ化はおこなわず、出現形が異なるものは別々に扱った。異なり数はおよそ 7 千であった。以後ことわりなく形態素という用語を用いた場合これらを指すものとする。

以下に分類される形態素を対象とした

助詞　助動詞	判定詞　接続詞	感動詞　副詞
形式名詞	時相名詞	副詞的名詞
名詞性述語接尾辞	形容詞性名詞接尾辞	形容詞性述語接尾辞
動詞性接尾辞		

活用しない形態素の例

格助詞　を	感動詞　じゃあ	副詞　さしずめ

活用する形態素の例

形容詞性名詞接尾辞	ダ列特殊連体形	がちの
形容詞性名詞接尾辞	ダ列基本推量形	がちだろう
形容詞性名詞接尾辞	ダ列基本条件形	がちならば
子音動詞カ行促音便形	基本形	く
子音動詞カ行促音便形	未然形	か
子音動詞カ行促音便形	意志形	こう
子音動詞カ行促音便形	省略意志形	こ
子音動詞カ行促音便形	命令形	け

2.3　指標の計算方法

Step 1. シードとなる会話文を選択する　はじめの教師となる会話文集合を定義する (以下シード)。WWW コーパス 3.7 億文のうち疑問の「か」を除く以下の終助詞を含む 0.3 億文をシードとした。

「かい」	「かしら」	「さ」	「さあ」	「ぜ」
「ぞ」	「っけ」	「で」	「な」	「なぁ」
「なあ」	「ね」	「ネ」	「ねぇ」	「ねえ」
「ねん」	「や」	「やろ」	「よ」	「ヨ」
「わ」				

Step 2. 形態素の生起確率をもとめる　各形態素について、会話文という条件の下での生起確率と特に条件を指定しないときの生起確率をもとめる。ユニグラムモデルを用いて最尤推定した場合、相対頻度を計数することに等しい。

ある形態素 w_1 の出現頻度を $f(w_1)$、生起確率を $p(w_1)$ とすると、会話文 (speech) から学習されたモデルとコーパス全体から学習されたモデルそれぞれでの 生起確率 (= 相対頻度) は

$$p(w_1|speech) = \frac{f(w_1|speech)}{\sum f(w_i|speech)} \quad (1)$$

$$p(w_1) = \frac{f(w_1)}{\sum f(w_i)} \quad (2)$$

である。

Step 3. 生起確率の比を指標とする　会話モデルでの生起確率と全体モデルでの生起確率の比をとり、その対数を形態素 w の口語性指標 i_w とした。

$$i_w = \log_2 \frac{p(w|speech)}{p(w)} \quad (3)$$

会話文でもそれ以外でも生起確率の変わらないニュートラルな形態素ならば 0 前後の値をとり、会話文で生起しやすい形態素は正の値、会話文で生起しにくい形態素は負の値をとる。

全体での頻度が 200 以上の形態素の指標のみ使用することとした。ゼロ頻度問題に対処するためのディスカウントを行った。全体で 200 回以上出現したにも関わらず会話文に出現しなかった形態素では

$$p(w|speech) = \frac{\sum f(w_i|speech)}{\sum f(w_i)} \quad (4)$$

とした。

Step 4. 文の会話文らしさを評価する　文中に出現した形態素の口語性指標 i_w を加算して、各文の会話文らしさ得点 I_s とした。学習できなかった (出現頻度が 200 未満であった) 形態素の i_w は 0 として計算した。

$$I_s = i_{w_1} + i_{w_2} + i_{w_3} + \cdots \quad (5)$$

これは、ユニグラムモデルにおける文生成確率の比を計算することと同等である。

Step 5. 値が収束するまで繰り返し計算をおこなう　$I_s > 0$ の文は全体モデル上よりも会話文モデル上で生成しやすいといえる。改めてこれを会話文集合とみなし、再び Step 2 - 4 を繰り返す。i_w が収束するまで繰り返した。

3　結果と考察

図 1 に i_w 値の収束する様子を示した。7 回目の計算がおわった時点で $I_s > 0$ と判定される文は 1.0 億文 (全体の 28%) であった。

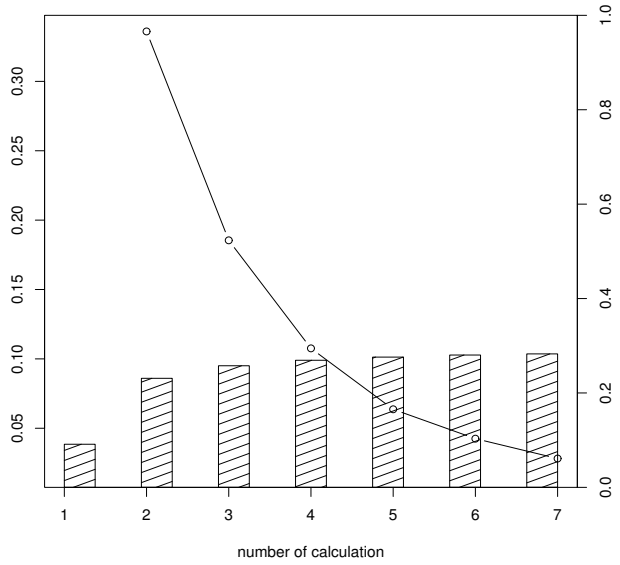

図 1: 繰り返し計算による値の収束。横軸は計算回数。折れ線 (左目盛) は再計算に伴う i_w の平均変動率 ($3 \times$ 四分位範囲(3IQR) に対する割合)。棒グラフ (右目盛) は会話文と判定される ($I_S > 0$ となる) 文の割合。

各形態素へ付与された指標値の、繰り返し計算にともなう平均変動率は単調に減少している。このまま収束するものと思われる。この手続きは文、または形態素を共起頻度を基準に 2 極へ分類するクラスタ分類とみなせる。i_w 系列の収束する先はコーパスの特性を反映した典型的な局所安定解に限られる可能性が高い。

各形態素に与えられた指標 i_w の例を示す (表 1)。正の値を付与された上位の形態素群は Cook(2001) [2] であげられているような音便形の語尾などをほぼカバーしている。affect keys の検出にはほぼ成功していると考える。また、終助詞と共起する文をシードとして出発したのだが、終助詞自体は最上位グループを形成したわけではなかった。降順ランクにして 100 位以内に入ったのは 90 位台の「なぁ」「ねぇ」のみであり、以下 100-800 位台にばらけ、$i_w = 0.17$ であった「や」などはほぼニュートラルな形態素といえる。これは得られた i_w 系列がシードよりもコーパスの特性に依存したものであることを示唆している。

学習コーパス内から $I_s > 0$ と評価された文例を示す (数字が得点)。

9.58 顔が「さわやか」してるから、嫌味にならないんだよな。

3.67 しかし、人数が多くて、ちょっと大変です。

8.34 軽い薬がだんだん効かなくなってきて、だんだん強い薬になってるんですよ

2.86 やっぱり、おかしいと思った。

次は $I_s < 0$ とされた文例である。

表 1: 形態素に付与された指標

値が正である形態素の例		
指標	分類	出現形
1.93	助動詞	ぱっかだ
1.88	形容詞活用語尾	けりゃ
1.87	終助詞	ねぇ
1.83	副詞	なんで
1.82	動詞活用語尾	ったろ
1.78	感動詞	へえ
1.77	接続助詞	けど
1.77	形容詞性述語接尾辞	めでしょう
1.72	形容詞性名詞接尾辞	っぽかった
1.70	副助詞	ったら
値が負である形態素の例		
指標	分類	出現形
-4.81	副助詞	など
-5.05	動詞活用語尾	じた
-5.36	感動詞	只今
-5.62	動詞性接尾辞	ございました
-7.28	接続詞	従って
-7.56	形容詞活用語尾	であったろう
-8.08	格助詞	にて
-8.21	助動詞	ばかりであり
-8.40	副詞	直ちに
-9.79	形容詞性述語接尾辞	がちであり

くだけた話しことば口調の形態素に大きな値、硬質な文章口調や敬語口調の形態素に小さな値が与えられている様子が観察できる。

-28.31 質問票は、具体的な個別問題に言及されており、多くは政府報告書の説明で足りない点について、情報を補足するよう促している。

-5.81 諸事情によりこのページは閉鎖しました。

-16.51 さらに料理の付け合わせにはフレッシュバジル、ミント、マナオなどがたっぷりついてきます。

-2.82 全寮制で半年間三ツ星ホテルに滞在する。

正と評価された文からは、確かに相手への何らかの働きかけを感じると言ってよさそうである。逆に情報を提示しているだけの文、何かを描写しているだけの文がマス体、ダ体の区別無く負と評価されている。

以上の観察結果から、指標 i_w がほぼ affection に対応し、得点 I_s がほぼ involvement に対応する尺度であるとみなしてよいと考えている。

構築した指標を使って現実の文を評価した例を示す (表 2)。(漫画) 欄の「これで元気に世にはばかってください」の評価は負である。明らかな会話文脈であっても常に high involved な発話ばかりが続くわけではなく、会話場が十分よく成立している限りにおいて、様々な幅をもった表現が容認されるという現実の会話状況をよく近似していると思われる。(学術講演) では、一見硬質な口調の中に含まれる若干くだけた表現をよく検出している。一文ごとに評価することで会話の展

表 2: 分野別、文の評価得点

得点	平均	小説の台詞: 坂口安吾『青鬼の褌を洗う女』(青空文庫 http://www.aozora.gr.jp/)
4.29	0.54	あなただって私をずいぶん悩ましたじゃないの
1.04	0.21	そして結局こんなふうになるわけか
3.51	0.88	罰が当るって、なによ
1.76	0.59	なんだい？　罰が当るって
2.50	0.83	いつか、あなた、いったでしょう。
5.73	0.57	オメカケが浮気してロクなことがあったタメシがないんだって。
3.61	0.90	罰が当るんだって。
-0.45	-0.11	そんなことをいったかしら。
		漫画: 二ノ宮知子『のだめカンタービレ』18 巻 (講談社 2007)
5.38	1.08	僕が死んじゃえなんて言ったから？
8.20	1.02	そんなの４０年前から言ってるじゃないですか！
2.05	0.34	「憎まれっ子世にはばかる」って言うのに
-3.20	-0.53	あなたが一番憎まれてますから
-2.60	-0.52	これで元気に世にはばかってください
2.99	0.60	さっ、もう出掛ける時間ですよ
2.48	0.50	いらない…今日は寝る…
0.90	0.23	今日はインタビューと撮影が！
		新聞記事: 毎日新聞 2002 年 11 月
-3.36	-0.67	同懇談会に台湾与党の民主進歩党から招へいがあった。
-6.79	-1.13	１７日に民進党と共催する第１回の台日政党シンポジウムに参加する。
-11.97	-1.33	現地では陳総統のほか、李登輝前総統との会談日程も調整している。
-11.65	-1.94	議長国のインド政府は同日夕、宣言の第２次案を公表。
-4.57	-0.91	住宅金融公庫は現行２・５５％の住宅ローン基準金利を２・４５％に引き下げる。
1.90	0.11	それなら、単に子供たちが、その活発な個性を発揮する方向が定まらずに苦しんでいるだけではないだろうか。
-2.21	-0.55	同党の地方組織で初めての試み。
		製品マニュアル: DR-2580C ユーザーズガイド (キヤノン 2006)
-10.01	-0.83	読み取りガラスにキズがあるとスキャンした画像にすじが入ったり、搬送エラーの原因になります。
-22.01	-2.00	市販の綿棒を使い、読み取りガラスの手前または奥にあるシェーディング板の汚れを拭き取ります。
-1.22	-0.24	上部ユニットを無理に閉じないでください。
-7.96	-1.59	本体の故障の原因になります。
-16.30	-2.33	搬送ローラーを交換後、以下の手順でカウンタをリセットしてください。
		学術講演: 日本語話し言葉コーパス (国立国語研究所 2004)
-15.00	-1.25	コウモリが目標物に到達する約秒前からパルスの送波が開始されることが確認できます
8.43	0.94	細かくてよく見えないかもしれないんですけれども
-4.87	-0.44	二個から五個パルスが一組みになってることが分かると思います
-23.68	-1.82	目標物に止まった場合で下が目標物の手前でターンした場合になっております
-6.91	-0.86	軸を見ていただいては右側の軸を見てください
-6.16	-0.77	同様の実験系で同様の実験を行なった結果ですけれども
-7.13	-1.19	目標物の前でターンした結果です

得点は一文あたり。平均は機能語句数で割った値。正の値に評価された文は会話文らしいと考えられる。

開を分析でき、生成に応用する際にも細やかな制御が可能になるであろうことを期待するものである。

4　おわりに

本稿では WWW コーパス上での機能語句の共起確率をもとに、各機能語句に情意性の強度を反映した数値指標を与えた。この指標が文体の話し言葉らしさをはかる尺度として利用できることを示した。本手法の特徴は一文ごとの口調評価が可能になる点である。現在、この評価尺度の文生成や言い換えシステムへの応用を検討中である。

参考文献

[1] N. Besnier. Involvement in linguistic practice: an ethnographic appraisal. *Journal of Pragmatics*, Vol. 22, pp. 279–299, 1994.

[2] H.M. Cook. An indexical analysis of the japanese naked plain form. 『日本語ディスコースへの多様なアプローチ』, pp. 73–99. 凡人社, 2007.

[3] 金明哲. 助詞の n-gram モデルに基づいた書き手の識別. 計量国語学, Vol. 23, No. 5, pp. 225–240, 2002.

[4] 泉子 K.Maynard. 『情意の言語学「場交渉論」と日本語表現のパトス』. くろしお出版, 2000.

[5] 鍛治伸裕, 岡本雅史, 黒橋禎夫. Www を用いた書き言葉特有語彙から話し言葉語彙への用言の言い換え. 自然言語処理, Vol. 11, No. 5, pp. 19–37, 2004.