文体横断的な文末機能の類似度測定

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

日本語の文では文末に述語がおかれる。その活用語尾や後接する接尾辞、助動詞などによって推量や意志といったモダリティ、命題の肯否など種々の機能を表現することができる。ところが、この文末形式には文体による多様性があって、たとえばフォーマルな文体とインフォーマルな文体では同等の文機能を実現するために違った形を用いることがある。

本稿では、文末表現の機能的類似性を文体の違いを超えて評価する方法を提案する。ここでいう文体とは時と場所、伝達媒体、受け手との関係性などによって規定される文章の様式のことである。文体のもうひとつの意味、書き手の個性を表す修辞的な傾向、は扱わない。

語句の意味を定量的に扱う手段として分布類似度という考え方がある [4]。「意味が類似した語は類似した文脈に出現する」という仮定に基づいて統計的な類似度評価をおこなうものである。いま、機能的な表現に対して分布類似度を定義しようと思う。機能的な表現は具体的な指示対象としての「意味」も、心的イメージとしての「意味」も有していないが、次のように考えれば概ね同様な議論が成立するであろう。すなわち、文は何がしかの認知的あるいは伝達的役割を果たすべく作文される。その際、文中の機能成分が連携して働くことによって特定の文機能が実現されている。よって、類似した機能を有する表現は同じような機能要素群と共起しがちである。具体的には、文末や文体と機能的な形態素の共起を考える。

2　資料と方法

本研究では以下の手順で構築した 14.5 億文のコーパスを用いた。WWW に由来する日本語文書 1 億ページ [2] から、河原, 他 (2006 [1]) と同様の手法により重複のない日本語文を抽出した。得られた 16 億文を形態素解析器 JUMAN、構文解析器 KNP で処理したのち、2 文節以上である文のみを取り出した。

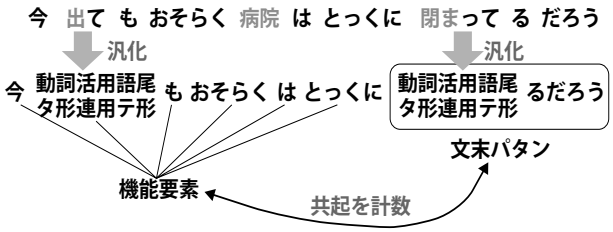

図 1: 文末パタンと機能要素の共起を考える

2.1　文末表現の類似度

文末表現の性質を記述するために機能的な形態素との共起尺度を並べた特徴ベクトルを定義し、ベクトル間の類似度を計算することで表現の類似性を評価する。

文末パタン　文末文節を取り出して、その構成形態素を以下の基準にしたがって汎化したものを文末パタンとよぶことにする (表 1)。文節単位の認定には構文解析器 KNP の文節まとめ上げ結果をそのまま用いた。

体言類 (名詞 名詞性述語接尾辞を除く名詞性接尾辞 連体詞) 連続する場合は一つにまとめた上で「名詞類」＋「最後の要素を品詞へ汎化したもの」の 2 要素として表現する。ただし文節頭の連体詞は汎化しない。

副詞類 (感動詞 副詞 時相名詞) 文節頭は汎化しない。それ以外は品詞へ汎化する。汎化された同じ品詞が連続する場合は一つにまとめる。

用言類 (動詞 形容詞) 活用語尾へ汎化する (品詞、活用型は区別する)。活用形が異なる場合別々に扱う。すなわちレンマ化はこれを行わない。

機能語類 (助詞 接続詞 形式名詞 名詞性述語接尾辞 助動詞 判定詞 形容詞性述語接尾辞 形容詞性名詞接尾辞 動詞性接尾辞) 汎化しない。レンマ化しない。

その他 (副詞的名詞 指示詞 接頭辞) 汎化しない。

表 1: 文末パタンの例

表現例	文末パタン
書いてました	[動詞活用語尾 いて] ました
よかったですよ	[イ形容詞活用語尾 かった] ですよ
素敵なんですよね	[ナ形容詞活用語尾 な] んですよね
もんだね	[形式名詞 もん] だね
広告会社じゃないですか	[名詞類, 普通名詞] じゃないですか
御確認願います	御 [サ変名詞][動詞活用語尾 い] ます

主辞と汎化要素に品詞を付した。実際は他の構成要素についても JUMAN の解析結果を保持している。

表 2: 機能要素の例

活用しないもの	
副詞	たぶん, おそらく, きっと
形式名詞	の, こと, わけ
接続詞	ただし, ゆえに, そして
活用するもの	
動詞活用語尾	たろう, けば, って
形容詞活用語尾	くて, じゃ, ならば
接尾辞	くださら, がちでしょ, っぽくて

品詞と出現形を示した。実際は品詞/活用型/活用形/出現形の組として各要素を識別する。

機能要素　形態素解析器 JUMAN の品詞体系を基準に以下を機能的な形態素群とみなし、まとめて機能要素とよぶことにする (表 2 へ例を挙げる)。活用するものはレンマ化せずに扱った。

助詞　接続詞　形式名詞　助動詞　判定詞　感動詞
副詞　時相名詞　副詞的名詞　指示詞　接頭辞　連体詞
名詞性述語接尾辞　形容詞性述語接尾辞　形容詞性名詞接尾辞
動詞性接尾辞　動詞活用語尾　形容詞活用語尾

共起尺度の計算　文末パタンと、文末以外の文節に出現した機能要素を組にして文内共起頻度を計数した。1 文内に同じ機能要素が複数回出現した場合は複数回数えた (図 1)。

共起尺度として自己相互情報量 (PMI) を使う。数えた共起のうち、文末パタン b との共起であったものの頻度を f_b、機能要素 w との共起であったものの頻度を f_w、b と w の共起頻度を $f_{b,w}$、すべての組の共起頻度の合計を f_{all} として推定すると

$$\mathrm{PMI}(b,w) = \log_2 \frac{p(b \cap w)}{p(b) \cdot p(w)} = \log_2 \frac{f_{b,w} \cdot f_{all}}{f_b \cdot f_w} \qquad (1)$$

である。実際には、低頻度要素が極端な値をとることを避けるために適当な補完量 c と、補正対数変換 $\log'$ を用いた補完自己相互情報量 (cPMI) の形で使用した。

$$\mathrm{cPMI}(b,w) = \log' \frac{f_{b,w} \cdot f_{all} + c}{f_b \cdot f_w + c} \qquad (2)$$

$\log'(x)$ は $x_{max} = f_{all}/f_b$ としたときに $\log'(0) = -\log'(x_{max})$ となるように均衡させた対数変換である。

$$\log'(x) = \log_2\{x + (1-x) \cdot \frac{r}{1-r}\}$$
$$(\text{ただし } c = \sqrt{f_b \cdot f_{all}}\,,\ r = f_b/f_{all}) \qquad (3)$$

特徴ベクトル　文末表現の性質を記述するものとして特徴ベクトルを定義する。コーパス中で 100 回以上出現した機能要素約 5,000 個を選択した。それに対応する cPMI を並べたものを特徴ベクトルとする。文末パタン b の特徴ベクトルは次のようになる。

$$\mathbf{v}_b = \{\, \mathrm{cPMI}(b,w_0), \ldots, \mathrm{cPMI}(b,w_n) \,\} \qquad (4)$$

類似度の計算　ふたつの文末パタン b_1 と b_2 が与えられた時に、文末ベクトル $\mathbf{v}_{b_1}$ と $\mathbf{v}_{b_2}$ がなす角のコサインを表現間の標準類似度 sim と定義する。

$$\mathrm{sim}(b_1,b_2) = \cos(\mathbf{v}_{b_1}, \mathbf{v}_{b_2}) = \frac{\mathbf{v}_{b_1} \cdot \mathbf{v}_{b_2}}{||\mathbf{v}_{b_1}|| \cdot ||\mathbf{v}_{b_2}||} \qquad (5)$$

2.2　特定文体文の抽出と文体ベクトル

コーパスからある文体、たとえば会話調に属している可能性の高い文を抽出する。抽出された文を有標文とよび、その集合を有標文コーパスとよぶ。最初に簡単な経験則に従って有標文を抽出する。会話調であれば終助詞の「ね」や「よ」などが含まれている文を選択する。この最初の有標文は特にシード文とよぶ。

ある機能要素のコーパス中での出現頻度を全機能要素の総頻度で割ったものをその機能要素の出現確率とする (確率を 0 にしないように適当に平滑化する)。文 l に含まれる機能要素の出現確率を掛け合わせたものを l の生成確率と定義する。コーパス全体から計算される生成確率を $p(l)$、有標文コーパスから計算される生成確率を $p_m(l)$ とする。

はじめに、コーパス全体における機能要素の出現頻度を計数して各文の生成確率 $p(l)$ を計算した。

繰り返し計算　以下の処理を繰り返し実行した。
有標文コーパス中に出現する機能要素の頻度を計数する。各文の生成確率 $p_m(l)$ を計算する。各文へ文体得点 $\mathrm{score}(l) = p_m(l)/p(l)$ を与える。得点上位の文を有標、下位の文を無標としてコーパスを再分割する。

条件設定　いくつかの条件を変更することで 3 つの異なった結果を得た (表 3)。それぞれ会話調、丁寧調、論説調とよぶことにする。会話調を構成するための条件は [3] によるものとほぼ同等である。予備実験の結果、与えるシード文の条件とともに文体得点順の上位何割を有標文として抽出するかという条件が特に重

表 3: コーパスから特定の文体を抽出した例

会話調
あ、スタッフで大瀧詠一のカレンダー持ってない人は買うようにね。
帰りはどんな顔して帰ってくるかなー。
まあ最初に舞台設定が京極堂シリーズと同じだということを断ってあるので
それもファンサービスということなんだろうが、これで初めて京極堂キャラ読む読者は可哀想だな。
丁寧調
私たちは、この番組を日本に紹介したいと願い、昨年秋から具体的に動き始めました。
気になる点がございましたら遠慮なくお問い合わせ下さい。
最後にフライターグは近日中に「ファッション掲示板」で特集したいと思いますので
お話が合うｊｏｕｒｎｅｙさんのガンガン書き込みお待ちしてます。
論説調
古代中国から伝わり、日本などの文化圏にも形を変えつつ影響したものと思われる。
癌患者はエジプトで化学療法などを受ける予定。
昨年の９月、海老名市内の交差点で信号を無視して二輪車に衝突し、男性に重傷を負わせたとして、
危険運転致傷等の罪に問われたH被告の公判が７月２５日に横浜地裁で開かれた。

要であった。その他の詳細については別の機会に譲りたい。

会話調 終助詞を手がかりにシード文を選択し、文体得点上位 30% を有標文として抽出した。

丁寧調 謙譲表現と尊敬表現を表す活用形を手がかりにシード文を選択し、文体得点上位 10% を有標文として抽出した。

論説調 「デアル体」の活用語尾を手がかりにシード文を選択し、文体得点上位 10% を有標文として抽出した。

文体ベクトル 文の文体が s であるという事象と、文に機能要素 w が出現するという事象の共起を考えることで、文末ベクトルの場合と同様の手順で文体 s の特徴ベクトルを構成することができる。有標として抽出された文の文体は s であると仮定して会話調、丁寧調、論説調の文体ベクトルを定義する。

$$\mathbf{v}_s = \{\, \mathrm{cPMI}(s, w_0), \ldots, \mathrm{cPMI}(s, w_n) \,\} \tag{6}$$

文末表現の文体尺度 文体ベクトル $\mathbf{v}_s$ と文末ベクトル $\mathbf{v}_b$ の間でもコサイン類似度を計算することができる。これを文末表現の、その文体らしさをあらわす文体尺度 $\mathrm{tone}_s(b)$ と定義する。

$$\mathrm{tone}_s(b) = \mathrm{sim}(s, b) = \cos(\mathbf{v}_s, \mathbf{v}_b) \tag{7}$$

2.3 文体成分の分離

単純な類似度計算では、文体の類似性も他の機能の類似性も一緒くたにして評価してしまうことになる。文末ベクトルから、文体ベクトル方向の成分を除くことで、文体の類似性による影響を低減する。

特徴ベクトル空間の中で会話調、丁寧調、論説調の 3 本の文体ベクトルが張る部分空間 V を考え、その直交補空間を $V^{\perp}$ とする。文末表現 b の特徴ベクトル $\mathbf{v}_b$ の V への正射影を $\mathbf{v}_b^S$、$V^{\perp}$ への正射影を $\mathbf{v}_b^c$ とすると次の関係が成立する。

$$\mathbf{v}_b = \mathbf{v}_b^S + \mathbf{v}_b^c \tag{8}$$

$\mathbf{v}_{b_1}$ と $\mathbf{v}_{b_2}$ の類似度の代わりに $\mathbf{v}_{b_1}^c$ と $\mathbf{v}_{b_2}^c$ の類似度を考え、脱文体類似度 $\mathrm{sim}^{\mathrm{c}}$ と定義する。

$$\mathrm{sim}^{\mathrm{c}}(b_1, b_2) = \cos(\mathbf{v}_{b_1}^c, \mathbf{v}_{b_2}^c) \tag{9}$$

3 結果と考察

文末パタンのうち最頻出 9 万パタンを対象として議論を進める。これはコーパスの 96.2% を被覆する。分布類似度は構成する形態素の品詞が異なるような表現を柔軟に扱うことができるが、見通しをよくするために主辞の品詞が一致する表現に限って検討する。例として「判断であろう」などの表現を汎化した「サ変名詞 であろう」というパタンを取り上げる。このパタンは論説調尺度の順位百分率にして上位 0.3% 目に位置する表現であり、強い論説調の表現である。

表 4 左が標準類似度最上位、右が脱文体類似度最上位である。これらは文体も機能も類似した表現群であると考えられる。脱文体類似度の値は標準類似度から比べて一貫して低下している。文体成分の分離で文体類似分の類似度が減少した結果であると考えられる。表中最も論説調尺度下位の「ではないだろうか」の評価が相対的に上昇しているのも文体の影響が軽減された結果だと考えられる。一方、両結果には概ね同じ表現が挙がっており、文体の影響が残っていることが示唆される。これは、文体ベクトルが各文体の最大公約数的なものであるため、類似度最上位群のような特徴ベクトルのスペクトルが極めて酷似した組み合わせに対しては効果が限定的であると解釈することができる。

さて、本稿の狙いは異文体間で機能が類似している表現を評価することであった。したがって、主に注目すべきは標準類似度が中程度であった組み合わせであ

表 4: 「サ変名詞 であろう」と類似した表現

標準類似度		会話調	論説調	脱文体類似度		会話調	論説調
0.744	すべきであろう	99.5	0.4	0.672	ではないだろうか	79.3	1.4
0.744	的である	99.0	0.2	0.659	すべきであろう	99.5	0.4
0.740	[動詞 よう]	99.2	0.4	0.653	[動詞 よう]	99.2	0.4
0.734	でもある	94.4	0.3	0.646	的である	99.0	0.2
0.732	しているのである	99.7	0.1	0.638	でもある	94.4	0.3
0.728	ではないだろうか	79.3	1.4	0.629	するであろう	99.8	0.5
0.723	されるのである	99.7	0.3	0.624	されているのである	99.5	0.3

主辞の「サ変名詞」を省略して記載。文体尺度は tone_s を 9 万パタン中での相対的な順位百分率 (パーセンタイル) に直したもの。その文体の中で上位何%目に位置しているかを示している。左:標準類似度最上位。右:脱文体類似度最上位。

表 5: 「サ変名詞 であろう」と類似し、かつ会話調である表現

会話調尺度上位 20%に制限				会話調尺度上位に制限 + $\text{sim}^c > \text{sim}$				
標準類似度		会話調	論説調	脱文体類似度		会話調	論説調	丁寧調
0.642	だろう	9.0	2.9	0.400	でしょうね	4.4	45.7	16.5
0.554	かもしれない	10.5	3.7	0.320	だろうね	10.0	39.8	68.9
0.518	ではある	19.1	9.2	0.317	だろうな	3.7	42.1	60.0
0.516	でしょう	6.1	14.7	0.314	なんだろうけど	9.1	42.8	66.9
0.511	なのだろうか	18.4	6.3	0.313	だろうが	8.4	45.0	78.9
0.487	だろうか	10.3	9.4	0.309	かと	1.3	57.8	45.6
0.431	かもしれません	10.7	12.7	0.309	なんだろう	2.9	39.4	66.1

主辞の「サ変名詞」を省略して記載。文体尺度は tone_s を 9 万パタン中での相対的な順位百分率 (パーセンタイル) に直したもの。左: 1. 対象を会話調尺度の上位 20%までに制限、2. 標準類似度順に並べた。右: 1. 会話調尺度の上位 20%までに制限、2. $\text{sim}^c > \text{sim}$ であったものを選択、3. 脱文体類似度順に並べた。

る。ここには文体が類似しているか、機能が類似しているか、あるいは両方が少し類似している表現が含まれている。その中から機能類似表現を選択的に検出、評価したい。ひとつの解決策は文体尺度によるフィルタリングである。すなわち必要に応じて「論説調でない」「会話調である」「丁寧調である」のように条件を制限した対象の類似度を比較する。表 5 左に「会話調である」条件によるフィルタリング例を示した。確かに会話調の文体にふさわしい表現が列挙されている。ただし、条件を満たしている中で論説調の強い表現を高く評価するという偏向があり、文体に中立な類似性評価とは言い難い。もうひとつの解決策が脱文体類似度である。類似度中程度の組み合わせでは、文体成分の分離によって文体類似表現の類似度が低下する一方で、機能類似表現の類似度は上昇する。$\text{sim}^c > \text{sim}$ であった表現に注目するとそれがわかる。表 5 右に基準を満たす表現を列挙した。論説調尺度と丁寧調尺度について中立に近い結果を示しており、文体の影響を抑えて類似性を評価している様子が見て取れる。

最後に、ここで検討した類似性評価手法は文体言い換え処理のための候補列挙法であると捉え直すことができる。表 5 右の結果から、丁寧調であるものを選択すれば若干丁寧な話し言葉、丁寧調でないものを選択すればくだけた話し言葉を構成することができる。この観点から処理結果をさらにいくつか挙げておく。

(妥当な) 判断であろう
→ 丁寧な会話： 判断でしょうね/なのですが/ですよね/なので...
→ くだけた会話： 判断だろうね/だろうな/なんだろうけど/だろうが...

(わかりやすく) 説明すべきだ
→ 丁寧な会話： 説明しないと/しなくては/しましょうね/してね...
→ くだけた会話：説明しろ/してほしいね/しろと/すればいいのに...

(琵琶湖の東西は) 二十二キロである
→ 会話： 二十二キロくらい/だ/なんです/ですね...

参考文献

[1] D. Kawahara and S. Kurohashi. Case Frame Compilation from the Web using High-Performance Computing. In *Proceedings of The 5th International Conference on Language Resources and Evaluation (LREC-06)*, pp. 1344–1347, 2006.

[2] K. Shinzato, T. Shibata, D. Kawahara, C. Hashimoto, and S. Kurohashi. TSUBAKI: An open search engine infrastructure for developing new information access methodology. *Proceedings of IJCNLP2008*, pp. 189–196, 2008.

[3] 玉城伸仁, 黒橋禎夫. 機能語句の話し言葉らしさ指標. 言語処理学会第 14 回年次大会, pp. 436–439, 2008.

[4] J. Weeds, D. Weir, and D. McCarthy. Characterising measures of lexical distributional similarity. 2004.