E-007

日本語の文の平均像を体現した文を探す (2) 平均からの距離

Finding the Average Sentences in Japanese: (2) Distance from the Avarage

近藤 秀 † Shu Kondo
佐藤 理史 ‡ Satoshi Sato
刀山 将大 † Masahiro Tachiyama
加納 隼人 ‡ Hayato Kanou

1 はじめに

日本語の文の平均像とはどのようなものであろうか。日本語初の大規模均衡コーパスである『現代日本語書き言葉均衡コーパス』(以下、BCCWJ) [1] の編纂により、この問いに答えることが可能となった。

本論文では、文献 [2] で抽出した各文の特徴量を用いて、平均像を体現した文を求める方法を検討する。具体的には、文を特徴ベクトルとして表現したとき、特徴ベクトルの集合に対して平均ベクトルを定義し、この平均ベクトルと各文の特徴ベクトル間に距離を定義する。この距離が最も小さい文を、平均像を体現した文とみなす。

2 使用する特徴量

文献 [2] で抽出した 6 グループ 17 種類の特徴量を表 1 に示す。平均像を体現した文の決定では、(a) 全ての特徴量を使う場合と、(b) アスタリスクを付与した 6 つの特徴量のみを使う場合の 2 種類を試す。後者は、6 グループの中からそれぞれ一つの特徴量を選んだものである。

3 平均からの距離

文献 [2] は、各特徴量において、その平均像を表す値として中央値が適切であると考察している。我々は、その考察に基づき、特徴量の値そのものではなく、順位に着目する。すなわち、ある特徴量のある値がその特徴量の値のリストの中で何位であるかを考える。

形式的には次のようになる。与えられた文集合 $\mathscr{S}$ において、各文 $S \in \mathscr{S}$ は n 次元の特徴ベクトルで表現できると仮定する。

$$S = \boldsymbol{v} = (v_1, v_2, ..., v_n) \tag{1}$$

ここで v_i は、整数または実数とする。

このベクトルを次のように順位ベクトル $\boldsymbol{r}$ に変換する。

$$\boldsymbol{r} = (r_1, r_2, ..., r_n) \tag{2}$$

$$r_i(v_i) = \frac{low_i(v_i) + high_i(v_i)}{2} \tag{3}$$

ここで low_i と $high_i$ は、それぞれ次の値を表す。

$$low_i(v_i) = \text{“特徴 } i \text{ における、値 } v_i \text{ の順位の下限”} \tag{4}$$

$$high_i(v_i) = \text{“特徴 } i \text{ における、値 } v_i \text{ の順位の上限”} \tag{5}$$

次に、文の理想的な平均像を表すベクトル $\boldsymbol{r_A}$ を次のように定義する。

$$\boldsymbol{r_A} = (r_M, r_M, ..., r_M) \tag{6}$$

$$r_M = \frac{1 + |\mathscr{S}|}{2} \tag{7}$$

ここで $|\mathscr{S}|$ は、文集合 $\mathscr{S}$ に含まれる文の数である。すなわち、r_M は順位の中央値である。

† 名古屋大学 工学部電気電子情報工学科
‡ 名古屋大学大学院 工学研究科 電子情報システム専攻

表 1: 文の特徴量

長さ (個数)	文字数＊
	短単位語数
	長単位語数
	文節数
読点 (個数)	読点＊
表記 (百分率)	ひらがな＊
	漢字
	その他
語種 (百分率)	和語＊
	漢語
	その他
品詞 (百分率)	助詞・助動詞＊
	名詞・代名詞
	動詞
	形状詞・連体詞・副詞・形容詞
	その他
難易度 (個数)	難しい語＊

最後に $\boldsymbol{r_A}$ と $\boldsymbol{r}$ の距離を次のように定義する。

$$dist(\mathbf{r}, \boldsymbol{r_A}) = \sqrt{\sum_{i=1}^{n} (r_i - r_M)^2} \tag{8}$$

この距離の小さなものほど平均像に近い文とみなす。なお、すべての特徴量がその特徴量の中央値と一致した場合でも、一般には、この距離は 0 とはならないことに注意されたい。

4 平均像を体現した文

文献 [2] は、文集合 $\mathscr{S}$ として BCCWJ の書籍に含まれる 534,240 文を採用し、各文の特徴量を計算した。まず、これらの値を式 (1) に従って、特徴ベクトル $\boldsymbol{v}$ に変換し、次に、式 (2–3) に用いて、順位ベクトル $\boldsymbol{r}$ に変換した。さらに、式 (6–7) により、文の平均像を表すベクトル $\boldsymbol{r_A}$ を定め、式 (8) を用いて、すべての文の順位ベクトル $\boldsymbol{r}$ に対して $dist(\boldsymbol{r}, \boldsymbol{r_A})$ を計算した。最後に、この $dist(\boldsymbol{r}, \boldsymbol{r_A})$ の値を用いて文集合 $\mathscr{S}$ をソートし、値の小さい順に順位を付与した。以上の処理を、すべての特徴量を使う場合と、6 つの特徴量のみを使う場合の、2 つの場合に対して行なった。

すべての特徴量を使う場合の距離、順位を $dist_a$、$rank_a$、6 つの特徴量のみを使う場合の距離、順位を $dist_6$、$rank_6$ と表す。表 2 に、$rank_a$ の 1 位〜 5 位の 5 文と $rank_6$ の 1 位〜 5 位の 5 文の各特徴量等を示す。この表の左端には、それぞれの特徴量の中央値とその値の順位 ((3) 式) を示した。なお、ここで取り上げた 10 文がそれぞれどのような文であるかを、表 3 に示した。

$rank_a$ の 1 位〜 5 位の 5 文の各特徴量は、おおよそ中央値と近い値となっている。これは、すべての特徴量を距離計算に用いているため、当然である。そのため、$rank_6$ でも、比較的順位が高い。これに対して、$rank_6$ の 1 位〜 5 位の 5 文は、距離計算に使用している 6 つの特徴量は、

表 2: 各文の特徴量

		中央値	順位	$rank_a$ 上位 A	B	C	D	E	$rank_6$ 上位 F	G	H	I	J
長さ	文字数＊	33	263018.5	27	36	36	31	37	33	33	33	35	35
	短単位語数	21	266851.5	22	24	23	22	22	22	22	22	22	22
	長単位語数	18	269451.5	19	21	20	18	21	17	20	18	21	15
	文節数	8	280000	7	8	9	7	9	8	9	8	9	6
	読点＊	1	250101	1	1	1	1	1	1	1	1	1	1
表記	ひらがな＊	54.5	265134.5	55.6	55.6	58.3	54.8	56.8	54.5	54.5	54.5	54.3	54.3
	漢字	28.6	269626	29.6	33.3	27.8	29.0	32.4	39.4	24.2	36.4	34.3	34.3
	その他	12.5	268176.5	14.8	11.1	13.9	16.1	10.8	6.1	21.2	9.1	11.4	11.4
語種	和語＊	65.0	266491	68.2	66.7	65.2	68.2	68.2	63.6	63.6	63.6	63.6	63.6
	漢語	15.4	264544	13.6	16.7	17.4	13.6	13.6	22.7	22.7	27.3	13.6	22.7
	その他	16.7	256668	18.2	16.7	17.4	18.2	18.2	13.6	13.6	9.1	22.7	13.6
品詞	助詞, 助動詞＊	37.0	268458	36.4	37.5	39.1	36.4	36.4	36.4	36.4	36.4	36.4	36.4
	名詞, 代名詞	27.6	268029	27.3	29.2	26.1	31.8	27.3	31.8	27.3	22.7	27.3	27.3
	動詞	12.5	265820	13.6	12.5	13.0	13.6	13.6	13.6	18.2	18.2	13.6	22.7
	形状詞, 連体詞, 副詞, 形容詞	4.3	269036	4.5	4.2	4.3	4.5	4.5	4.5	0.0	4.5	4.5	0.0
	その他	15.6	267846	18.2	16.7	17.4	13.6	18.2	13.6	18.2	18.2	18.2	13.6
	難易度＊	2	305133.5	2	1	1	2	1	2	2	2	2	2
$dist_a$ ($\times 10^4$)		4.54		14.18	14.55	15.12	15.58	17.00	30.20	33.04	36.88	18.97	35.99
$rank_a$		-		1	2	3	4	5	2176	5056	13661	21	11066
$dist_6$ ($\times 10^4$)		4.19		9.24	8.69	11.15	6.85	1.15	4.86	4.86	4.86	5.09	5.09
$rank_6$		-		684	506	1911	140	1179	1	1	1	4	4

順位の中央値は、267120.5 である。

表 3: 文 A–H

A	これは私の最後の生である！」という、歓喜の宣言をした。
B	友だちからの刺激，物（人形）を媒介にして子どもは大きく変化したのである。
C	芸風にも共通したところがあるが、二人とも日本人には受けるタイプのようだ。
D	佐渡では長さ四〇尺の竹を、五〇〇円で売っていることもあります。
E	それで、夫婦で給料を持ち帰る仕事に就くことを「共稼ぎ」と呼びだしたようだ。
F	党内に大型補正への抵抗があることについて、森派幹部はこう洩らした。
G	６　ろくろを回しながら、最初の１本をドベを塗った上に積み始めます。
H	そうすれば、保証金と成約預託金の一、二割を返してやってもかまわん。
I	恐ろしい火事を「江戸の華」と呼ぶところに、皮肉な詩情が込められている。
J	２　倉荷証券については、預かり証券についての規定を読みかえて適用する。

中央値と近い値であるが、それ以外の特徴量の値は、必ずも中央値と近い値とはなっていない。たとえば、文 H では、表記において、ひらがなの割合は中央値に近いが、漢字やその他の割合は、中央値から離れている。さらに、語種において、和語の割合は中央値に近いが、漢語とその他の割合は中央値から離れている。これらの要因により、文 H の $rank_a$ は 13,661 位となっている。

同じグループに属する複数の特徴量の値は、相互に関係する。たとえば、語種のグループでは、和語の割合が増加すれば、一般に、漢語の割合が減少する。しかし、その他 (外来語や混種語など) が存在するため、和語の割合が同じであっても、漢語の割合が異なる場合がある。これらのことを考慮し、平均像を体現した文の決定では、多少冗長であっても、すべての特徴量を用いた方がよいと考える。

以上の議論に基づくと、今回の調査の帰結である、日本語の文の平均像を最もよく体現した 5 文は、文 A–E となる。

中央値の $dist_a$ と $dist_6$ は、対象とするすべての特徴量が中央値と一致した場合の距離の値、すなわち、距離の最小値を示している。$rank_6$ 上位の $dist_6$ の値はこの最小値に近いが、$rank_a$ の上位の $dist_a$ の値は、最小値にそれほど近くない。つまり、17 種類の特徴量がすべて中央値とほぼ一致するような文は、存在しなかったということである。

本研究では、BCCWJ の解析済データから比較的簡単に取り出せる特徴量 (17 種類) を用いた。当然のことながら、用いる特徴量のセットが異なれば、結果は異なる。今回は、品詞の割合を 5 種類の特徴量にまとめたが、この点については、より詳細な検討が必要である。同時に、より高次の特徴量 (たとえば、文の構造を反映した特徴量) なども検討する必要がある。

謝辞　本研究では、国立国語研究所編纂の『現代日本語書き言葉均衡コーパス』を使用した。本研究は、JSPS 科研費 24300052 の助成を受けた。

参考文献

[1] 国立国語研究所. 現代日本語書き言葉均衡コーパス. http://www.ninjal.ac.jp/corpus_center/bccwj/ (2014.6.29 にアクセス).

[2] 刀山将大, 佐藤理史, 近藤秀, 吉田達平. 日本語の文の平均像を体現した文を探す (1) 文の特徴量の抽出. FIT2014, 情報処理学会・電子情報通信学会, 2014.