E-006

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

Finding the Average Sentences in Japanese: (1) Feature Extraction

刀山 将大 † 佐藤 理史 ‡ 近藤 秀 † 吉田 達平 ‡
Masahiro Tachiyama Satoshi Sato Shu Kondo Tappei Yoshida

1 はじめに

日本語の文の平均像を体現した文というものは、どのような文であろうか。そのような文は存在するのであろうか。均衡コーパスは、この問いに答えるために必要なデータを提供する。本研究では、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) の解析済データを利用し、この問いに答えることに取り組んだ。

その第一段階として、まず、調査対象とする文の集合を定め、それらの各文の文字長、長単位語数、短単位語数、文節数、品詞分布などの特徴量を調査した。本論文では、その方法と結果について報告する。

2 調査対象と方法

2.1 調査対象サンプル

BCCWJ のサンプルには、1,000 字で構成される固定長サンプルと、章や節などの単位 (10,000 字以内) で構成される可変長サンプルの、2 つの形式が存在する。本研究では、統計処理を目的として設計された固定長サンプルを対象とする。

BCCWJ の固定長サンプルのレジスタは、出版-書籍、雑誌、新聞、図書館-書籍、白書の 5 種類である。このうち、本研究では、出版-書籍と図書館-書籍のみを対象とする。すなわち、本研究では、書籍に掲載されたテキストを日本語の母集団と仮定する。

2.2 調査対象文

本研究では、DVD 版に収録された BCCWJ の解析済データ (TSV データ) を利用する。このデータは、テキストを文分割したのち、語分割 (形態素解析) した結果が格納されている。ここでの「文」は、文区切り記号 (「。」や「.」など) を末尾に持つ、いわゆる普通の文以外に、タイトル (見出し) や箇条書の番号などの疑似的な文が含まれる。そのため、調査対象に、これらの疑似的な文を含めるかどうかの選択肢が存在する。本研究では、疑似的な文を含める場合 (Include) と除外する場合 (Exclude) の両者の場合について、調査を行なう。

2.3 調査方法

BCCWJ は、長単位語と短単位語という 2 種類の言語単位を採用している [1]。特徴量の算出には、主に短単位語を使用する。

特徴量の算出は、次の手順で行なう。

1. BCCWJ の短単位解析済データ (TSV データ) から、固定長フラグが 1 となっている文 (短単位語) を抜き出す。但し、文頭の全角スペースのみで構成される短単位語 (段落の字下げ) は除外する。
2. それぞれの文に対して、その文が疑似的な文であるかどうかを、文末に文区切り記号 (「。」「.」「？」「！」) の有無で判定する。
3. それぞれの文に対して、あらかじめ定めた 17 種類の特徴量のうち、長単位語数、文節数を除く 15 種類を計算する。

† 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: 調査対象

	Include	Exclude	
疑似的な文	含む	含まない	
サンプル数	20,668	20,668	
文数	689,987	534,240	(77.4%)
短単位語数	15,112,834	13,444,968	(89.0%)
長単位語数	12,606,540	11,220,662	(89.0%)

なお、特徴量の残りの 2 つである長単位語数と文節数は、長単位解析済データに対し、同様の手順を適用して計算する。

表 1 に、調査対象の文数、短単位語数、長単位語数を示す。この表より、疑似的文の数は全体の文数の 22.6%、それらの文に含まれる短単位語数と長単位語数は、ともに 11.0%であることがわかる。

17 種類の特徴量 (表 2) は、大きく、長さ、読点、表記、語種、品詞、難易度の 6 グループに分類される。これらのグループのうち、表記は文中の文字数に対する百分率、語種と品詞は文中の短単位語数に対する百分率で表す。品詞は、出現率が低い品詞が存在するため、似たような品詞をまとめ、5 種類に分類した。

難易度は、その文に出現する難しい語の個数である。本研究では、難しい語を、20,668 サンプル中 105 サンプル未満 (0.5%未満) にしか出現しない語と定めた。なお、難しい語に分類されない平易語は、異なりで 5,776 語である。

3 調査結果

調査対象の各文に対し 17 種類の特徴量を計算し、それらの最小値、最頻値、平均値、中央値、最大値を求めた。その結果を表 2 に示す。なお、百分率の最頻値は、1%刻みで計算した。

3.1 疑似的な文の影響

疑似的な文を含めた場合 (Include) と除外した場合 (Exclude) を比較すると、長さ、特に、文字数や短単位語数において、差が顕著である。図 1 に、文長 (文字数) とその文長をとる文の数の関係を示す。このグラフの x 軸は、対数スケールである。このグラフから、特に、文字数が少ないところで、Include と Exclude の差が大きいことがわかる。これは、疑似的文の文長 (文字数) が短いことに起因する。すなわち、Include では、この影響を大きく受けることになる。

その一方で、特徴量が百分率として計算する 3 つのグループの特徴量 (表記、語種、品詞) では、Include と Exclude の差は小さい。これは、これらの特徴量が長さで正規化されているためと考えられる。

表 2: 特徴量の統計量

分類	特徴量の種類	Include					Exclude				
		最小	最頻	中央	平均	最大	最小	最頻	中央	平均	最大
長さ	文字数	1*	14	28	35.1	6093	1*	23	33	40.2	6093
(個数)	短単位語数	1	10	18	22.0	3285	1*	16	21	25.2	3285
	長単位語数	1	9	15	18.3	2923	1*	13	18	21.0	2923
	文節数	0*	3	6	7.9	1432	1	6	8	9.0	1432
読点 (個数)		0	0	1	1.2	63	0	1	1	1.4	61
表記	ひらがな	0	0	52.9	49.4	100.0	0	50	54.5	53.3	97.6
(割合)	漢字	0	0	28.0	29.7	100.0	0	33	28.6	29.1	97.2
	その他	0	7	14.0	21.0	100.0	1.0	7	12.5	17.6	100.0
語種	和語	0	0	62.8	58.0	100.0	0	67	65.0	62.8	97.1
(割合)	漢語	0	0	15.4	18.5	100.0	0	0	15.4	17.2	91.7
	その他	0	13	18.2	23.5	100.0	1.2	13	16.7	20.1	100.0
品詞	助詞・助動詞	0	0	35.7	32.4	100.0	0	33	37.0	35.7	85.7
(割合)	名詞・代名詞	0	33	28.1	30.8	100.0	0	33	27.6	28.0	100.0
	動詞	0	0	11.5	11.2	100.0	0	0	12.5	12.5	66.7
	形状詞・連体詞・副詞・形容詞	0	0	3.3	5.1	100.0	0	0	4.3	5.5	90.9
	その他	0	13	16.7	20.3	100.0	0	13	15.6	18.2	100.0
難易度 (難しい短単位語の個数)		0	0	1	2.0	301	0	1	2	2.3	301

注：固定長サンプルは 1,000 字であるが、サンプル部分を少しでも含む文は、その文全体を調査対象とした。このため、文字長は 1,000 字を超える場合がある。なお、'*' を付与した数字は、TSV データの誤りと思われる。

我々が持つ、「文」の標準的なイメージは、末尾に述語を持ち、句点で終る文である。そのような文の平均像を体現した文を探すのであれば、疑似的な文を除外した文集合を対象とするのが適切と考えられる。

3.2 平均像の基準

特徴量の平均像を表す値として、算術平均、中央値、最頻値の 3 つの候補がある。値の分布が正規分布であれば、これらの 3 つの値は一致するが、一般には一致しない。平均像を体現した文を探す場合、これらのうち、どの値を基準値として採用すべきかを決定する必要がある。

表 2 に示した特徴量のうち、個数を値とする 3 つのグループ (長さ、読点数、難易度) では、3 つの値の大小関係は、

$$\text{最頻値} < \text{中央値} < \text{算術平均} \quad (1)$$

となる。これは、図 1 のように、値の大きい側がロングテールとなっているからである (図 1 の x 軸は対数スケールである)。このような分布の場合、算術平均は、極度に大きな値を持つサンプルの影響を大きく受けるため、平均像を表す基準値として適切ではない。残された候補は、最頻値と中央値であるが、最頻値を採用した場合は、算術平均の場合と逆に、値の大きい側がロングテールとなっていることを全く反映しないこととなる。

一方、百分率を値とするグループ (表記、語種、品詞) では、算術平均と中央値にそれほど大きな差はみられない。しかし、最頻値は、出現率の低い特徴量 (漢語、動詞、形状詞・連体詞・副詞・形容詞) で 0%となる。このため、平均像を表す基準値として適切ではない。

以上の検討に基づき、本研究では、各特徴量の平均像を表す基準値として、中央値を採用するのが適切であると考える。

4 まとめ

本論文では、日本語の文の平均像を体現した文を探すための第一段階、すなわち、対象とする文集合の選定と各文の特徴量の抽出について報告した。実際に平均像を体現した文を探す第二段階は、文献 [2] で報告する。

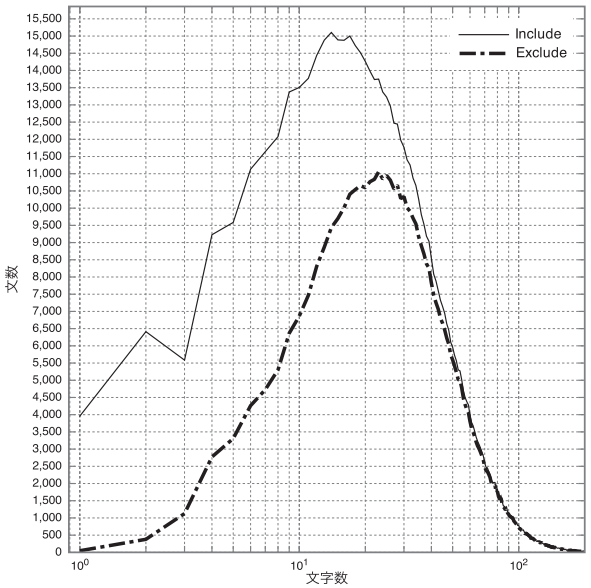

図 1: 文字数と文数の関係

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] 小椋秀樹, 小磯花絵, 冨士池優美, 宮内佐夜香, 小西光, 原裕. 『現代日本語書き言葉均衡コーパス』形態論情報規定集第 4 版 (上). 国立国語研究所内部報告書 LR-CCG-20-05-01, 国立国語研究所, 2011.

[2] 近藤秀, 佐藤理史, 刀山将大, 加納 隼人. 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. FIT2014, 情報処理学会・電子通信学会, 2014.