劣勢な表記を検出する作文支援システム

西川 彩　西村 涼　渡辺 靖彦　岡田 至弘
龍谷大学 理工学部 情報メディア学科
t060606@mail.ryukoku.ac.jp, r_nishimura@afc.ryukoku.ac.jp, {watanabe,okada}@rins.ryukoku.ac.jp

1　はじめに

日本語の文書では、1 つの単語に 2 つ以上の表記が存在する表記のゆれがよく見られる。表記のゆれは、情報検索や形態素解析の問題としてよく研究されているが [1] [2] [3] [4]、作文支援における問題としては、同一文書内での表記の一貫性の保持以外、これまであまり取り上げられていない。しかし、表記のゆれがある場合、どの表記がその文書において好ましいのか書き手が判断するのを支援することは重要である。例えば、(例文 1) と (例文 2) は下線部の表記だけが異なる文であるが、

(例文 1) なぜ 恐竜は絶滅したのですか。

(例文 2) 何故 恐竜は絶滅したのですか。

(例文 1) ではなく (例文 2) が文章中に用いられていれば、なぜわざわざ「なぜ」のかわりに「何故」を用いたのだろうかという疑問と奇妙な印象を読み手に与えかねない。これは、「なぜ」に比べて「何故」が劣勢な表記であるからである。劣勢な表記は誤りではなく、その使用は制限されるべきものではない。しかし、特に目的や理由がないのであるなら、書き手の不利益にならないように、できるだけ優勢な表記を使用することがのぞましい。それでも、あえて劣勢な表記を使用する場合は、

- 劣勢な表記を使用していることを書き手が認識していること
- 劣勢な表記をあえて使用する目的や理由が書き手にあること

が重要になる。しかし、大学の授業で提出されるレポートなどを調べると、劣勢な表記を使用していることを書き手が認識していないことが多い。この原因の 1 つは、どの表記が優勢/劣勢であるのか判断することがむずかしいことである。例えば、新聞記事における植物の名前の表記では、

- ひらがなによる表記が優勢なもの (ひまわり、みかん)
- カタカナによる表記が優勢なもの (バラ、リンゴ)
- 漢字による表記が優勢なもの (桜、椿)

があり、それぞれの植物の名前についてどの表記が優勢なのかを判断するのは簡単ではない。さらに、文書の種類によって優勢な表記が変化することがある。例えば、新聞記事では「桜」が優勢な表記であるが、植物図鑑の説明文では「サクラ」が優勢な表記である。

そこで本研究では、劣勢な表記が文書内で用いられていることを文書の作成者に知らせ、その表記の利用の目的や理由について再検討する機会を与える作文支援システムを作成することを目的とする。検出した劣勢な表記を優勢な表記に自動的に置換しないのは、

- 劣勢な表記であってもその使用を制限するべきではない
- 劣勢な表記を使用していることを書き手に認識させ、その目的や理由について検討する機会を与えることは、特に教育機関では意味がある

と考えたからである。作成したシステムでは、以下の方法で作文支援を行う。まず、ユーザが入力した文章に対して形態素解析を行い、その結果から表記のゆれがある単語を検出する。そして、その単語の表記が劣勢な表記ではないか、表記のゆれの辞書を利用して確認する。表記のゆれの辞書は、新聞記事と専門文書を用いて作成する。劣勢な表記である場合は、その単語の表記のゆれの頻度情報をユーザに示し、表記の選択について再検討することを促す。

2　新聞記事と論文における表記のゆれ

本研究では、劣勢な表記を検出するのに、

- 新聞記事
- 専門文書

における表記のゆれの情報を用いる。新聞記事からは、特定の分野に限らず広く用いられる単語の表記のゆれの情報を収集する。一方、専門文書からは、書き手が作成しようとしている文書の分野特有の単語の表記のゆれの情報を収集する。表記のゆれの情報を取り出す専門文書を変更すれば、文書の種類によって変化する優勢/劣勢な表記にも対応できる。本研究では、新聞記事および専門文書における表記のゆれの情報を以下の文書を対象に調査して収集した。

- 毎日新聞一年分 (2006 年) の 296364 記事
- 言語処理学会の年次大会 (2006 年) に投稿された 319 個の論文

表 1 に、新聞記事および論文に含まれる単語で、表記のゆれがあると判定されたものの表記の異なりと出現頻度を示す。表記のゆれがある単語かどうかは、JUMAN[4] を用いて形態素解析した結果得られる代表表記を用いて判定した。JUMAN[4] によって表記のゆれがあると判定された単語 (新聞記事:27988 語、論文:9211 語) は、以下の 2 つに分類できる。

- 新聞記事/論文では表記が 1 つだけ検出された単語 (新聞記事:19406 語、論文:7849 語)
- 新聞記事/論文で複数の表記が検出された単語 (新聞記事:8582 語、論文:1362 語)

表 1　新聞記事 [毎日 2006] と論文 [言語処理学会 2006] に含まれる単語で、表記のゆれがあると判定されたものの表記の異なりと出現頻度

品詞	単語の異なり	表記の異なり	表記の出現頻度
名詞	20603	26747	3656574
動詞	3897	6403	1283024
形容詞	2120	2830	280787
副詞	1125	1607	115609
接続詞	87	100	30850
感動詞	80	97	2643
連体詞	75	98	10946
接頭辞	1	3	10891
合計	27988	37885	5391324

(a)　新聞記事 [毎日 2006] に含まれる単語で、表記のゆれがあると判定されたものの表記の異なりと出現頻度

品詞	単語の異なり	表記の異なり	表記の出現頻度
名詞	6458	7154	310980
動詞	1548	2093	101398
形容詞	706	825	22952
副詞	376	459	13037
接続詞	60	71	4465
感動詞	30	33	148
連体詞	32	39	1192
接頭辞	1	3	302
合計	9211	10677	454474

(b)　論文 [言語処理学会 2006] に含まれる単語で、表記のゆれがあると判定されたものの表記の異なりと出現頻度

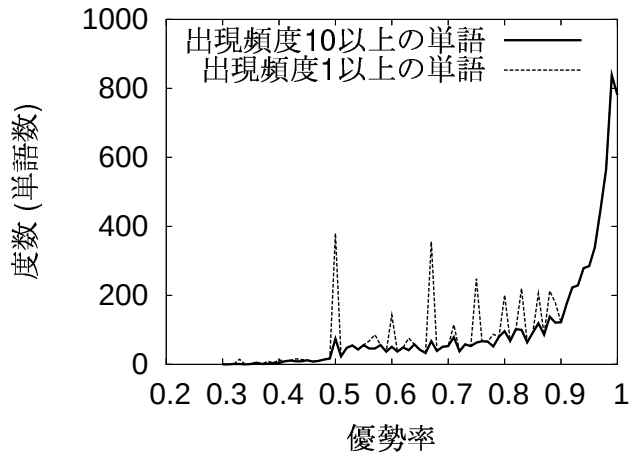
表 2　新聞記事 [毎日 2006] と論文 [言語処理学会 2006] で複数の表記が検出された単語の表記の異なりと出現頻度

品詞	単語の異なり	表記の異なり	表記の出現頻度
名詞	5328	11472	1817055
動詞	2135	4641	916302
形容詞	628	1338	176374
副詞	440	922	72251
接続詞	13	26	12980
感動詞	15	32	593
連体詞	22	45	8853
接頭辞	1	3	10891
合計	8582	18479	3015299

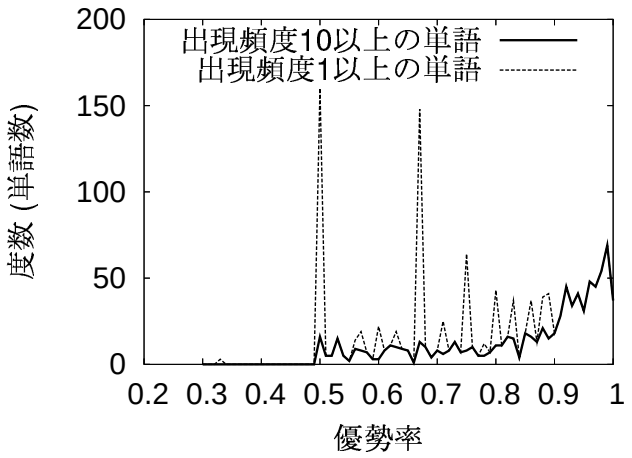
(a)　新聞記事 [毎日 2006] で複数の表記が検出された単語の表記の異なりと出現頻度

品詞	単語の異なり	表記の異なり	表記の出現頻度
名詞	644	1340	62848
動詞	508	1053	56058
形容詞	110	229	6253
副詞	78	161	5617
接続詞	11	22	1330
感動詞	3	6	13
連体詞	7	14	941
接頭辞	1	3	302
合計	1362	2828	133362

(b)　論文 [言語処理学会 2006] で複数の表記が検出された単語の表記の異なりと出現頻度



(a)　新聞記事 [毎日新聞 2006] で複数の表記が検出された単語の優勢率のヒストグラム



(b)　論文 [言語処理年次大会 2006] で複数の表記が検出された単語の優勢率のヒストグラム

図 1　新聞記事 [毎日 2006] および論文 [言語処理年次大会 2006] で複数の表記が検出された単語の優勢率のヒストグラム

表 2 に、新聞記事/論文で複数の表記が検出された単語の表記の異なりと出現頻度を示す。新聞記事/論文で複数の表記が検出された単語について、最も優勢な表記がどれぐらい優勢に用いられているのかを評価するため、以下のように優勢率を定義する。

$$\text{優勢率} = \frac{\text{対象となる語で最も優勢な表記の出現頻度}}{\text{対象となる語のすべての表記の総出現頻度}}$$

図 1 に、新聞記事/論文で複数の表記が検出された単語の優勢率に関する以下の 2 種類のヒストグラムを示す。

- 新聞記事/論文で複数の表記が検出された単語 (新聞記事:8582 語、論文 1362 語) の優勢率のヒストグラム (図 1 の破線)
- 新聞記事/論文で複数の表記が検出され、すべての表記の総出現頻度が 10 以上の単語 (新聞記事:6949 語、論文:819 語) の優勢率のヒストグラム (図 1 の実線)

すべての表記の総出現頻度が 10 未満の単語を取りのぞいて優勢率のヒストグラムを作成したのは、頻度 10 未満の単語については新聞記事/論文の用例だけでは優勢/劣勢の判定に失敗するおそれがあると考えたからである。したがって、表 2 に示す新聞記事/論文で複数の表記が検出された単語 (新聞記事:27988 語、論文:9211 語) の内、優勢な表記についての比較的信頼できる情報を取り出せたと考えられるものの内訳を以下に示す。

- 新聞記事/論文で表記が 1 つだけ検出された単語 (優勢率は 100%) で、その出現頻度が 10 以上のものは、新聞記事で 11825 語、論文で 2285 語あった。
- 新聞記事/論文で複数の表記が検出された単語で、すべての表記の総出現頻度が 10 以上、優勢率が 80% 以上のものは、新聞記事で 5270 語、論文で 590 語あった。優勢率が 60% 以上のものは、新聞記事で 6347 語、論文で 744 語あった。

3 劣勢な表記を検出する作文支援システム

作成した作文支援システムの概要を図 2 に示す。図 2 に示すように、システムは以下の 3 つのモジュールから構成されている。

インターフェース インターフェイスには、Web ブラウザを用いた。ユーザは、Web ブラウザを利用してシステムに文章を入力し、劣勢な表記の検出結果を受け取る。

劣勢表記検出モジュール 劣勢表記検出モジュールは、JUMAN[4] を用いて形態素解析を行い、その結果得られる代表表記を手がかりにして表記のゆれのある単語を検出し、その表記が劣勢な表記ではないか、表記のゆれの辞書を利用して判定する。劣勢な表記である場合は、その単語の表記のゆれについての頻度情報も表記のゆれの辞書から取り出し、劣勢な表記の検出結果としてインターフェイスに送る。

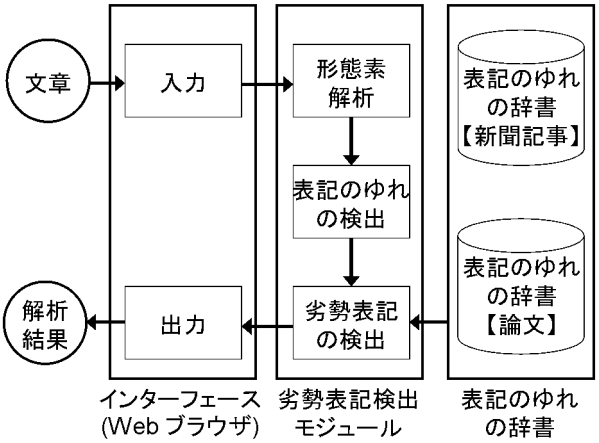


図 2 劣勢な表記を検出する作文支援システムの概要

表記のゆれの辞書 表記のゆれの辞書は、2 章で述べた

- 新聞記事 [毎日 2006] から取り出した表記のゆれのある単語 (27988 語) とその表記 (37885 表記) の頻度情報
- 論文 [言語処理学会 2006] から取り出した表記のゆれのある単語 (9211 語) とその表記 (10677 表記) の頻度情報

から構成されている。劣勢表記検出モジュールからの照会に応じて、単語の表記が優勢であるか劣勢であるか判定する。劣勢な表記であった場合、その単語の表記のゆれの頻度情報も劣勢表記検出モジュールに送る。

4 表記のゆれの頻度情報を用いた表記の選択の実験

作成した作文支援システムでは、劣勢な表記を検出すると、ユーザにその単語の表記のゆれの頻度情報を示して表記の選択について再検討を促す。そこで、表記のゆれの頻度情報が表記の選択に有効かどうか確認するために実験を行った。

実験では、単語の表記が 1 つだけ異なる 2 つの文を 10 組用意し、大学の授業のレポートで利用するのにのぞましいと思う表記を含む文を被験者に選択させた。実際に実験に用いた問題を図 3 に示す。被験者は情報系の大学生 (2 回生)20 人で、10 人ずつ以下の 2 つのグループにわけて実験を行った。

グループ A 図 3 の問題のみが与えられている。

グループ B 図 3 だけでなく、表記のゆれの頻度情報も与えられている。例えば、図 3 の (問 4) ならば、以下に示す表記のゆれの頻度情報が与えられる。

	むずかしい	難しい
新聞記事	21	1524
論文	0	155

(問 1) a. なぜ あなたはこの本を書き始めたのですか？
b. 何故 あなたはこの本を書き始めたのですか？
(問 2) a. たとえば、2004 年には約 2.1 兆円の輸出をしている。
b. 例えば、2004 年には約 2.1 兆円の輸出をしている。
(問 3) a. ただし、TF-IDF 法だけを用いて索引語を抽出する。
b. 但し、TF-IDF 法だけを用いて索引語を抽出する。
(問 4) a. たばこをやめるのは むずかしい。
b. たばこをやめるのは 難しい。
(問 5) a. たばこをやめるのは やさしい。
b. たばこをやめるのは 易しい。
(問 6) a. 東ドイツ政府は、西ドイツへ自由に出国することができる という誤報を流した。
b. 東ドイツ政府は、西ドイツへ自由に出国することが出来る という誤報を流した。
(問 7) a. 考察を 行なう 必要がある。
b. 考察を 行う 必要がある。
(問 8) a. 動物園の象は野生の象よりも寿命が 短かい。
b. 動物園の象は野生の象よりも寿命が 短い。
(問 9) a. わき水の透明度は 変わる ことがある。
b. わき水の透明度は 変る ことがある。
(問 10) a. 文中の文法構造を 表わす。
b. 文中の文法構造を 表す。

図 3　表記の選択実験に用いた問題文 (実験では、表記のゆれがある単語以外も読ませるため、下線は引いていない)

グループ	表記選択	問題 1	2	3	4	5	6	7	8	9	10
A	a	4	2	8	2	6	5	2	1	10	2
	b	6	8	2	8	4	5	8	9	0	8
B	a	7	0	6	0	9	5	0	0	10	0
	b	3	10	4	10	1	5	10	10	0	10

図 4　図 3 の問題に対する表記の選択実験の結果 (グループ A: 頻度情報なし、グループ B: 頻度情報あり、下線部: 優勢な表記を選択した人数)

図 4 にグループ A と B の表記の選択結果を示す。さらに、グループ A と B の

- 優勢な表記の選択率
- グループ内での表記の選択の一致 (κ 値)

を表 3 に示す。表 4 には Landis らによる κ 値の解釈を示す [5]。表 3 の κ 値が示すように、グループ A では表記の選択はあまり一致していない。一方、グループ B では表記の選択がかなり一致している。さらに、優勢な表記の選択率もグループ A に比べて 13% 向上している。これらは、表記のゆれの頻度情報が表記の選択に有効であることを示している。

グループ B の被験者に聞き取り調査を行ったところ、6 人の被験者からは表記のゆれについての頻度情報が表記の選択に有効であったとの回答を得た。具体的には、

- 選択しようとしていた表記が劣勢な表記であることに表記のゆれについての頻度情報によって気づき、選択を優勢な表記に変更した (3 人)
- 優勢な表記を選択しようとしていることを頻度情報によって知り、その選択に確信がもてた (3 人)

また、表記のゆれの頻度情報という具体的な根拠を示されるので、表記の選択の変更も素直にできるとの意見もあった。一方、残りの 4 人からは頻度情報は必要ないとの回答を得た。具体的には、

- 頻度情報がなくても優勢な表記を選択できる (1 人)
- 大学のレポートではできるだけ漢字を使用するのがのぞましいので、頻度情報は参考にせず、漢字による表記を選択した (3 人)

表 3　グループ A と B の優勢な表記の選択率とグループ内での表記の選択の一致 (κ 値)

グループ	優勢な表記の選択率	κ 値
A	74%	0.261
B	87%	0.623

表 4　κ 値の解釈 (Landis ら [5] による)

κ	Interpretation
< 0	no agreement
0.0 - 0.20	slight agreement
0.21 - 0.40	fair agreement
0.41 - 0.60	moderate agreement
0.61 - 0.80	substantial agreement
0.81 - 1.00	almost perfect agreement

謝辞 本研究の一部は、日本学術振興会科学研究費補助金基盤 (C)「心豊かなコミュニケーションを促進する質問作成支援システムの作成」(課題番号 20500106) の助成を受けて行われたものです。

参考文献

[1] 久保村, 亀田: 片仮名異表記処理能力を備えもつ情報検索システム, 電子情報通信学会論文誌, Vol.J86-D-II, No.3, (2003).
[2] 甲田: 科学技術文献検索システムにおける異表記対応について, 情報処理学会研究報告, 2006-FI-85, (2006).
[3] 馬場, 新里, 黒橋: 検索エンジン基盤 TSUBAKI を用いた大規模ウェブ情報クラスタリングシステムの構築, 情報処理学会研究報告, 2008-NL-183, (2008).
[4] 黒橋, 河原: 日本語形態素解析システム JUMAN version 5.1 使用説明書, 京都大学, (2005).
[5] Landis and Koch: The measurement of observer agreement for categorical data, Biometrics, Vol. 33, (1977).