係り受け関係を考慮した優勢表記辞書の作成

西川 彩　渡辺 靖彦　西村 涼　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, watanabe@rins.ryukoku.ac.jp, r_nishimura@afc.ryukoku.ac.jp,
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 [3] [4] [5] [7]、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望ましい。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難しいことにある。表記の選択は重要な問題であるが、これまであまり研究されていない。横山は異体字の選択について報告しているが [6]、単語の表記の選択については扱っていない。

そこで、この表記選択の問題を解決するために、劣勢な表記を検出して優勢な表記をユーザに提示する作文支援システムを作成した [1]。このシステムは、新聞記事や専門文書で最も優勢に用いられる表記 (以後、優勢表記とよぶ) が好ましい表記であるという仮定に基づいている。ある表記が優勢表記であるかどうかは、新聞記事や専門文書における表記のゆれを調査して作成した優勢表記辞書を利用して判定する。しかし、この優勢表記辞書の作成では係り受け関係を考慮していないため、以下の問題があった。

- 単語の意味の解釈にあいまいさがあり、優勢表記の判定に利用できない用例が数多くある
- 特定の係り受け関係があると、例外的に優勢に用いられる劣勢表記がある

そこでこれらの問題を解決するため、係り受け関係を考慮した優勢表記辞書を作成する方法について述べる。

2　係り受け関係を考慮しなければ扱えない表記のゆれ

作文支援システム [1] で用いる優勢表記辞書を作成するために、新聞記事および専門文書から表記のゆれがある語を取り出し、その頻度情報からどの表記が優勢表記であるかを推定した。表記のゆれがある語は、JUMAN[7] 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検出した。しかし、JUMAN[7] による形態素解析では意味の解釈にあいまいさがある例が数多くあり、それらは優勢表記の推定には利用できなかった。例えば、(例文 3) の「おかす」は JUMAN[7] による形態素解析では意味の解釈にあいまいさがある例で、3 つの代表表記 (「侵す」「犯す」「冒す」) が与えられていて、そのままでは優勢表記の推定には利用できない。

(例文 3) リスクを おかす。

しかし、こうした例の中には、異表記の係り受け関係を考慮することで解釈のあいまいさを解消することができるものがある。例えば、(例文 4) の「冒す」は (例文 3) の「おかす」の場合と同様に「リスク」と係り受け関係がある。

(例文 4) リスクを 冒す。

表 1 に「おかす」「侵す」「犯す」「冒す」の新聞記事 (2006 年 1〜6 月) における出現頻度を示す。表 2 には、「リスク」と係り受け関係がある「おかす」「侵す」「犯す」「冒す」の出現頻度を示す。表 2 に示すように、「冒す」は「侵す」や「犯す」に比べ、「リスク」との係り受け関係がある例が多い。そこで、(例文 3) の「おかす」の意味は (例文 4) の「冒す」と同じであると考えられる。

次に、係り受け関係を考慮しなければ扱えない表記のゆれの例を示す。

(例文 5) 成功の カギ はあなたが握っている。

(例文 6) 家の 鍵 をかけるのを忘れた。

表 3 に示すように、新聞記事新聞記事 (2006 年 1〜6 月) における「かぎ」の優勢表記は「カギ」である。しかし、表 4 に示すように、「かぎ」が「かける」と係り受け関係にある (例文 6) のような場合、「鍵」が例外的に優勢に用いられる。このように、特定の係り受け関係では例外的に

表 1　「おかす」の表記のゆれの出現頻度 (毎日新聞 [2006 年 1〜6 月])

おかす	侵す	犯す	冒す
4	73	127	16

表 2　「リスク」が係る「おかす」とその異表記の出現頻度 (毎日新聞 [2005〜2007 年])

おかす	侵す	犯す	冒す
1	0	0	19

表 3　「かぎ」「カギ」「鍵」の出現頻度 (毎日新聞 [2006 年 1〜6 月])

かぎ	カギ	鍵
1	279	198

表 4　「かける」に係る「かぎ」とその異表記の出現頻度 (毎日新聞 [2005〜2007 年])

かぎ	カギ	鍵
0	10	64

優勢に用いられる劣勢表記の情報を優勢表記辞書に登録しないと、その文脈では劣勢な表記を優勢な表記であるとユーザに誤解させるおそれがある。したがって、特定の係り受け関係では例外的に優勢に利用される劣勢表記について調べる必要がある。

3　係り受け関係を考慮した優勢表記辞書

係り受け関係を考慮した優勢表記辞書は、係り受け関係を考慮しないで作成した優勢表記辞書 [2] に係り受け関係を考慮した優勢表記の情報を追加することによって作成する。係り受け関係を考慮しないで作成した優勢表記辞書は、毎日新聞に 2006 年 1 月〜6 月に掲載された 296364 記事を用いて作成した。この辞書には 20929 語の優勢表記が登録されているが、それらの語は以下に示す TYPE I と TYPE II の 2 つのタイプに分けられる。

TYPE I　表記が 1 つだけ検出された単語 (14659 語)
TYPE II　複数の表記が検出された単語 (6270 語)

特に TYPE II の語には、特定の係り受け関係がある場合に例外的に優勢に用いられる劣勢表記をもつ単語が含まれている。この TYPE II の語で優勢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、優勢率を以下のように定義する。

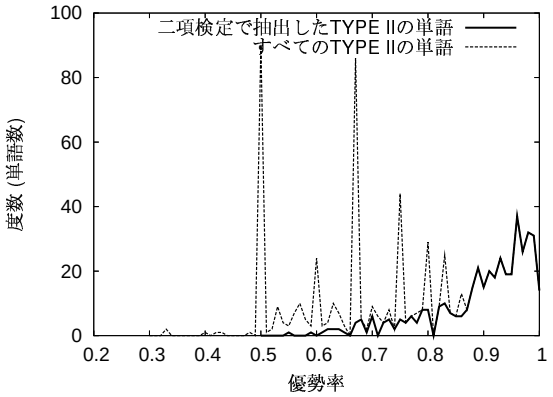

図 1　TYPE II の単語の優勢率のヒストグラム

ある単語が複数の表記をもっているとする。そのうち、表記 i $(\in I)$ の利用率 u_i を

$$u_i = \frac{f_i}{\sum_{i \in I} f_i}$$

と表す。f_i は表記 i の頻度である。このとき、単語の優勢率 d は

$$d = \max_{i \in I} u_i$$

と表すことができる。

この優勢率に着目し、優勢表記が変化するおそれの少ない単語、すわなち、信頼できる優勢表記を二項検定を用いて取り出した。図 1 に、新聞記事で用いられていた TYPE II の単語の優勢率のヒストグラムを示す。点線は、TYPE II のすべての単語の優勢率のヒストグラムを示している。一方、太線は、二項検定で抽出した TYPE II の単語の優勢率のヒストグラムを示している。このようにして作成した優勢表記辞書に係り受け関係を考慮した優勢表記の情報を追加するため、本章では以下の方法について述べる。

- 異表記の係り受け関係の情報を用いて、意味解釈のあいまいさを解消する方法
- 特定の係り受け関係がある場合、例外的に優勢に用いられる劣勢表記を抽出する方法

3.1　係り受け関係を用いた意味解釈のあいまいさの解消

異表記の係り受け関係を用いて意味解釈のあいまいさを解消する方法を以下に示す。

step 1　新聞記事を対象に形態素解析および係り受け解析を行う。形態素解析には JUMAN [7]、係り受け解析には KNP [8] を用いた。

step 2　step 1 の結果から、JUMAN によって複数の代表表記を与えられた単語、すなわち、意味解釈があいまいな単語を含む係り受け関係を取り出す。

step 3 step 2 で取り出した係り受け関係に含まれる意味解釈があいまいな単語 (例:(例文 3) の「おかす」) について、意味解釈にあいまいさがない異表記 (例:(例文 4) の「冒す」) で同じ係り受け関係をもつものを取り出す。また、その頻度情報も取り出す。

step 4 step 3 で取り出した異表記を含む係り受け関係の中で、最も出現頻度が高い係り受け関係に含まれる異表記の意味を意味解釈があいまいな単語の意味とする。

step 5 step 4 の意味解釈が信頼できるかどうかを二項検定を用いて判定する。具体的には、最も優勢な解釈の出現率の片側 95% 信頼区間の下限が 0.5 よりも大きい単語を信頼できる意味解釈として抽出した。

この方法によって、毎日新聞に 2005 〜 2007 年に掲載された 1786752 記事 [9] に含まれる 34037 個の係り受けに含まれる単語の意味解釈のあいまいさを解消できた。

3.2 特定の係り受け関係で優勢表記が変わるもの抽出

特定の係り受け関係がある場合、例外的に優勢に用いられる劣勢表記を以下の方法で取り出す。

step 1 新聞記事を対象に形態素解析および係り受け解析を行なう。形態素解析には JUMAN [7]、係り受け解析には KNP [8] を用いた。

step 2 step 1 の結果から、表記のゆれがある単語を含む係り受け関係を取り出す。さらに、それらの頻度情報も取り出す。ただし、3.1 節で述べた方法で意味解釈のあいまいさを解消できない単語を含むものは取り出さない。

step 3 step 2 で取り出した係り受け関係の中から、優勢表記辞書 [2] には劣勢表記として登録されているが、その係り受け関係では優勢に用いられる表記を含むものを取り出す。

step 4 step 3 で取り出した係り受け関係に含まれる劣勢表記が、その係り受け関係では優勢表記として信頼できるかどうかを二項検定を用いて判定する。具体的には、その出現率の片側 95% 信頼区間の下限が 0.5 よりも大きい場合、その係り受け関係では信頼できる優勢表記として取り出した。

この方法によって、毎日新聞に 2005 〜 2007 年に掲載された 1786752 記事 [9] に含まれる 3598 個の係り受け関係について、例外的に優勢に用いられる劣勢表記を取り出した。

4 表記のゆれの頻度情報を用いた表記選択実験

作成した辞書の有効性を評価するため、大学生を対象に表記の選択について対照実験を行った。実験では、下線が引かれた 1 つの単語を含む 2 つの文で構成された問題を 5 組用意し、大学の授業のレポートなどで利用するのに望ましいと思う表記を選択肢から選択させた。それぞれの問題の最初の文は、係り受け関係を考慮していない優勢表記辞書 [2] では優勢表記であるとされている表記が優勢に用いられている例文である。一方、2 番目の文は、その係り受け関係から優勢表記辞書 [2] では劣勢表記であるとされている表記が優勢に用いられている例文である。実際に実験で用いた問題を図 2 に示す。例えば、問題 (1-a) の「ひきあげる」では、優勢表記辞書 [2] で優勢表記とされている「引き上げる」が優勢に用いられている。一方、「投資」が「ひきあげる」に係っている問題 (1-b) の場合、劣勢表記である「引き揚げる」が優勢に用いられている。被験者は情報系の大学生 (3 回生) 20 人で、10 人ずつ、対照群と実験群の 2 つのグループに分けて実験を行った。図 3 に実験の概要を示す。図 3 に示すように、対照群にはテスト 1 とテスト 2、実験群にはテスト 1 とテスト 3 を順に与えた。それぞれのテストで被験者に与える情報を以下に示す。

テスト 1 情報なし
テスト 2 係り受け関係を考慮していない優勢表記の情報
テスト 3 係り受け関係を考慮した優勢表記の情報

例えば、テスト 2 では、問題 (1-a) と (1-b) の両方で、表 5 に示す頻度情報を被験者に与える。ただし、この情報は問題 (1-b) の場合、劣勢表記を優勢表記であると被験者を誤解させるおそれがある。一方、テスト 3 では、問題 (1-a) ではテスト 2 と同様に表 5 に示す頻度情報を被験者に与えるが、問題 (1-b) では、係り受け関係を考慮して、表 6 の頻度情報を与える。

表 7 にテスト 1、2、3 での優勢表記の選択率を示す。表 7 は、表記選択が重要な問題であることを示している。テスト 1 では、劣勢表記をそうであるとは気づかずに、あるいは特に理由なく、選択している被験者がいた。一方、テスト 2 、3 では、テスト 1 での表記の選択にこだわりをもたず、根拠や理由が与えられれば柔軟に選択を変える被験者がいた。実際、テスト 3 では、5 人の被験者が選択を変えている。さらに、他の 2 人は選択を変えていないが、与えられた情報によって自分の選択に確信が持てたと述べている。また、教員の「こういう表記は使用しないこと」という指示よりも、表記のゆれの頻度情報という具体的な根拠を示す方法の方が表記の選択を素直に変更できるとの意見もあった。一方、テスト 2 でも、5 人の被験者が選択を変えている。しかし、そのうち 2 人は、係り受け関係を考慮していない優勢表記の情報のため、その係り受け関係では劣勢な表記を選択していた。これらの結果から、係り受け関係を考慮しない優勢表記の情報でユーザを混乱させ

(1–a) 閣議で税金を ひきあげる ことが決定された。
1. ひきあげる　2. 引きあげる
3. 引き上げる　4. 引き揚げる
(1–b) ニューヨークの市場から投資を ひきあげる ことにした。
1. ひきあげる　2. 引きあげる
3. 引き上げる　4. 引き揚げる
(2–a) 中学の頃の友人と偶然 であう。
1. であう　2. 出あう　3. 出逢う
4. 出会う　5. 出合う
(2–b) 図書館で素敵な文学作品に であう。
1. であう　2. 出あう　3. 出逢う
4. 出会う　5. 出合う
(3–a) 成功の かぎ はあなたが握っている。
1. かぎ　2. カギ　3. 鍵
(3–b) 家の かぎ をかけるのを忘れた。
1. かぎ　2. カギ　3. 鍵
(4–a) 敵と たたかう ゲームが好き。
1. たたかう　2. 戦う　3. 闘う
(4–b) 病気と たたかう 決心をした。
1. たたかう　2. 戦う　3. 闘う
(5–a) 答えを答案用紙に うつす。
1. うつす　2. 映す　3. 写す　4. 移す
(5–b) 自分の姿を鏡に うつす。
1. うつす　2. 映す　3. 写す　4. 移す

図 2　表記の選択実験に用いた問題文

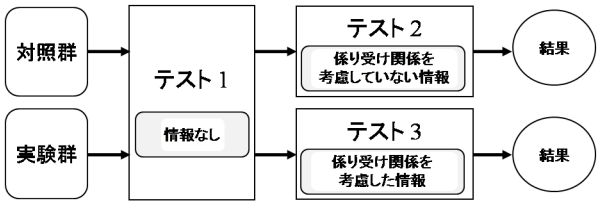

図 3　実験の概要

ないため、表記選択を支援する作文支援システムでは、係り受け関係を考慮した優勢表記辞書を用いることが重要であることが分かった。

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] 西川, 西村, 渡辺, 岡田: 劣勢な表記を検出する作文支援システム, 言語処理学会第 15 回年次大会, P3-5, (2009).
[2] 西川, 渡辺, 西村, 村田, 岡田: 表記選択支援のための優勢表記辞書の作成, 電子情報通信学会技術研究報告, NLC2009-3, (2009).
[3] 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).
[4] 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).
[5] 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).
[6] 横山: 字体選好は新聞頻度から予測可能か, 情報処理学会研究報告. 人文科学とコンピュータ研究会報告, (2006).
[7] 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).
[8] 黒橋, 河原: 日本語構文解析システム KNP version 2.0 使用説明書, 京都大学, (2005).
[9] 毎日新聞 CD-ROM データセット 2005, 2006, 2007, 日外アソシエーツ株式会社, (2006-2008).

表 5　「ひきあげる」の表記のゆれの出現頻度 (毎日新聞 [2006 年 1〜6 月])

ひきあげる	引きあげる	引き上げる	引き揚げる
1	4	774	146

表 6　「投資」が係る「ひきあげる」とその異表記の出現頻度 (毎日新聞 [2005〜2007 年])

ひきあげる	引きあげる	引き上げる	引き揚げる
0	0	2	15

表 7　係り受け関係を考慮した優勢表記の選択率

グループ	テスト 1	テスト 2/3
対照群	68%	77%
実験群	73%	81%