機械学習を用いた表記選択の難易度推定

小島 正裕† 村田 真樹‡ 南口 卓哉† 渡辺 靖彦†
†龍谷大学大学院 理工学研究科 情報メディア学専攻
{t10m101, t060629}@mail.ryukoku.ac.jp, watanabe@rins.ryukoku.ac.jp
‡鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
murata@ike.tottri-u.ac.jp

1 はじめに

日本語の文書では、同じ単語が異なった表記で用いられる表記のゆれがよく見られる。日本語は同じ単語でも漢字表記、仮名表記、片仮名表記という表記の違いがある言語である。また、送り仮名の付け方によっても表記が変わったり、「醤油」を「しょう油」と書くなど、漢語の一部を仮名で書く交ぜ書きによる表記の違いもある。このような表記のゆれがある単語を文中で用いる場合、どの表記を用いるかは判断に迷うことが多い。例えば、「病気とたたかう」と書きたい場合、「戦う」なのか「闘う」なのかについて判断に悩む。このような場合、1 つの解決策として辞書などを頼りにすることが挙げられるが、辞書を参照してもその区別は不明瞭であることが多い。そのため、これまでに、単語の頻度情報をもとに表記選択の検討を促す研究 [5] や、新聞コーパスとアンケート調査をもとにした表記の使い分けの考察 [7] が行われている。表記のゆれがある単語について、どのような単語が表記選択が容易で、どのような単語が表記選択が困難か、その傾向や特徴を捉えられれば、今後の表記のゆれの研究やシステムに役立つと考えた。

そこで、本研究では、機械学習で表記選択が容易であるか否かに基づいて、人間による表記選択の難易度を推定する。例えば、機械学習による表記選択が困難であったものは人間による表記選択も困難であると予想する。一方、機械学習による表記選択が容易であったものは人間による表記選択も容易であると予想する。この手法により、各単語について表記選択が難しいものかそうでないものかを明らかにし、表記の選択が容易な場合はなぜ容易であったかを明らかにする。2 節で、本研究の立場および考え方を述べる。次に、3 節で、実験対象のデータについて説明する。最後に、4 節で、実験結果と考察を行う。

2 本研究の立場・考え方

これまでに表記のゆれは、情報検索や曖昧性解消の問題としてよく研究されている [1, 2, 6]。また、本研究の立場でもある表記の選択や使い分けといった観点での研究もされている。西川らは、新聞・論文に使用されている表記のゆれがある単語の頻度情報を用いて、頻度の最も高いものを優勢な表記、それ以外のものを劣勢な表記としている。そして、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研究を行っている [5]。日本らは、「かえる (変える・替える・換える・代える)」という単語について、名詞との共起関係に着目し、新聞コーパスとアンケート調査をもとにした表記の使い分けの考察をしている [7]。

本研究では、表記のゆれがある単語について機械学習による表記選択を行う。機械学習は、素性と呼ばれる情報を利用して表記選択を行うので、頻度情報だけではわからない名詞や用言との共起関係による表記の使い分けがある場合に、正しい表記選択をすることができると考えた。機械学習によって高い正解率で表記選択を行えたものは人間による表記選択が容易で、機械学習によって高い正解率で表記選択を行えなかったものは人間による表記選択が困難である可能性が考えられる。どのような単語が表記選択が容易で、どのような単語が表記選択が困難か、その傾向や特徴を捉えられれば、今後の表記のゆれの研究やシステムにも役立つと考えた。

本研究では、表記のゆれがある単語を、機械学習の正解率の高さごとに、高・中・低の 3 つに分類する。どのような基準で分類するかは 4 節で詳しく述べる。高い正解率で表記選択を行えたものと高い正解率で表記選択を行えなかったものについて、なぜそのような結果になったのかを考察する。そして、機械学習が正しく判定できたものは人間でも正しく判定できるのか、機械学習が正しく判定できなかったものは人間でも正しく判定できないのかを調査する。また、機械学習の正解率の高さごとに何か傾向や特徴があるかを考察する。本研究では、機械学習法として最大エントロピー法 (以下、MEM) を用いる。

3 機械学習を用いた表記選択

本研究では、表記のゆれがある単語について機械学習を用いて適切な表記を判定し、その結果を考察する。

本節では、最初に、実験に利用するデータについて説明する。次に、機械学習に与える素性について説明する。

3.1 実験で用いるデータ

本研究では、機械学習を用いた表記選択をするのに、新聞記事における表記のゆれの情報を用いる。新聞記事における表記のゆれの情報は、毎日新聞（2005〜2007 年）の 3,693,567 文を対象に調査して収集した。新聞記事に含まれる単語で、表記のゆれがあると判定されたものは 29,815 語あり、そのうち表記が 1 つだけ検出された単語は 14,630 語、複数の表記が検出された単語は 15,185 語あった。表記のゆれがある単語かどうかは、JUMAN[3] 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。複数の表記が検出された 15,185 語に対し、以下の条件でデータを抽出する。

(条件 1) 対象の単語のすべての表記の合計出現頻度数が 100 以上であるもの

(条件 2) 対象の単語の曖昧性を避けるため、JUMAN の解析結果で@マークが一度もつかないもの

(条件 3) 対象の単語の各表記の出現頻度数上位 2 つが、どちらも 10 以上であるもの

(条件 1) は、新聞記事内でよく使われている単語について調査を行うためである。(条件 2) の「JUMAN の解析結果で@マークがつかないもの」とは、表記は違うが代表表記が同じものである。逆に、JUMAN の解析結果で@マークのつくものは、代表表記が別の単語であることを示している。例えば、「けいじ」という単語を JUMAN で解析すると代表表記が「啓示」のほかに、@マークがつき代表表記が「掲示」「計時」「刑事」が解析結果として出力される。「啓示」「掲示」「計時」「刑事」はそれぞれ別の単語である。このように、JUMAN の解析では、読みは同じで代表表記が別の単語がある場合は、先頭に@マークをつけて出力する。このような単語を避けるために (条件 2) をつけた。表記選択の実験は、その単語の各表記の出現頻度数上位 2 つで行うのだが、どちらか一方の出現頻度数が低すぎると機械学習を行う上で不具合が生じるため、(条件 3) をつけた。これらの条件と一致する単語数は 1,877 語であった。本研究では、この 1,877 語のうち無作為に取り出した 939 語を実験対象とする。

3.2 素性

機械学習を用いた表記選択を行うために用いる素性を表 1 に示す。この素性は表記選択をしたい表記のゆれがある単語を持つ文から取り出す。素性によって機械学習は、その文においてどちらの表記が適切であるかを判定する。また、表 1 における分類番号とは、分類語彙表に記されている語の意味ごとに与えられる 10 桁の番号のことである [9]。単語の中でも意味が複数あるものにはその数だけ番号がふられている。本研究では、番号を 5 桁、3 桁に区切り素性として与えている。これにより、それぞれの単語の上位概念を素性とすることができる。

表記選択を行うためには、表記のゆれがある単語が含まれている文において、表記のゆれがある単語の前後の情報が有効だと考えた。そこで、s1-s20 の素性を定義した。これらの素性は、表記選択を行う対象となる表記のゆれがある単語を含む文節に関する情報を利用している。文の構造を明らかにする係り受けの情報を、表記選択の素性として利用することは有効だと考えた。そこで、s21-s60 の素性を定義した。これらの素性は、対象となる表記のゆれがある単語の係り元や係り先に、表記選択に役立つ情報がある場合に有効だと考えられる。さらに、表記選択を行う対象となる表記のゆれがある単語の前後にある文字列が、表記選択に有効だと考え s61、s62 を定義した。

本研究では、係り受け関係を素性として扱うために、構文・格解析システム KNP[4] を用いる。

4 実験と考察

4.1 実験方法

機械学習は、実験データとして抽出した 939 語について、各単語ごとに適用した。各単語ごとに、10 分割のクロスバリデーションを行った。機械学習は、表記のゆれがある単語の各表記の出現頻度数上位 2 つについて判定を行う。

4.2 実験結果

再現率の高さごとの傾向や特徴を捉えるために、機械学習の再現率の高さごとに高・中・低の 3 つに分類する。再現率の高さごとに分類するのは、再現率が実験データにある正解データのうち、機械学習がどれだけ正解を認識したかを示す割合であるため、正解率と同義だからである。2 つの表記のうち、低いほうの再現率が 80%より高い単語を再現率高とする。これは、両方の再現率が適度に高い値でないと適切な判定ができていないと考えたからである。2 つの表記のうち、低いほうの再現率が 50%より高く 80%以下である単語を再現率中とする。2 つの表記のうち、低いほうの再現率が 50%以下である単語を再現率低とする。これは、片方の再現率が高くても、もう片方の再現率が極端に低いものは全体としては再現率の高い結果でないと考えたからである。このように、機械学習の再現率の高さごとに高・中・低の 3 つに分類した結果は表 2 のようになった。

表 2 より、機械学習による表記選択の実験としては、表記のゆれがある単語 939 語中 81 語について、それぞれの単語の各表記の出現頻度数上位 2 つの表記とも 80%以上の再現率で適切な表記を選択できたこととなる。本研究の主たる目的は、機械学習によって正しく表記選択を行えるものは人間でも正しく表記選択が行えることを確認することではあるが、副次的な効果として、本研究の機械学習は、表記選択自体にも役立つことがわかる。

4.3 考察

4.3.1 単語ごとの考察

機械学習の再現率の高さごとに表記選択を行った単語をいくつか抜粋し、各単語の表記選択の難易度がどの程度であったかを考察する。各単語ごとに、例文、機械学習の結果、機械学習が判定の参考にした素性をそれぞれ示す。例文は、機械が判定した結果の正解例（機械が例文と同じ表記を選んだ例）と失敗例（機械が例文と異なる表記を選んだ例）を 2 文ずつ示す。機械学習が判定の参考にした素性は、正規化 α 値の高いものから 3 つを示す。正規化 α 値とは、素性ごとに、MEM で求まる α 値を全分類先での和が 1 になるように正規化したものである。ある分類先の正規化 α 値が高い素性ほど、その分類先であることを推定するのに役立つ素性である。分類先 a と素性 s の正規化 α 値が x である場合、素性 s だけで判断する場合、分類先 a である確率は x であることを意味する [8, 10]。

表 1: 多義性の解消に用いる素性

番号	素性の種類
s1	表記選択を行う対象の単語を含む文節の最初の自立語
s2	s1 の品詞
s3	s1 の自立語の分類語彙表での分類番号 5 桁までの数字
s4	s1 の自立語の分類語彙表での分類番号 3 桁までの数字
s5	表記選択を行う対象の単語を含む文節の最後の自立語
s6	s5 の品詞
s7	s5 の自立語の分類語彙表での分類番号 5 桁までの数字
s8	s5 の自立語の分類語彙表での分類番号 3 桁までの数字
s9	表記選択を行う対象の単語を含む文節の自立語
s10	s9 の品詞
s11	s9 の自立語の分類語彙表での分類番号 5 桁までの数字
s12	s9 の自立語の分類語彙表での分類番号 3 桁までの数字
s13	表記選択を行う対象の単語を含む文節の最初の付属語
s14	s13 の品詞
s15	表記選択を行う対象の単語を含む文節の最後の付属語
s16	s15 の品詞
s17	表記選択を行う対象の単語を含む文節の付属語
s18	s17 の品詞
s19	表記選択を行う対象の単語を含む文節の記号
s20	s19 の品詞
s21	表記選択を行う対象の単語を含む文節に係る文節の最初の自立語
s22	s21 の品詞
s23	s21 の自立語の分類語彙表での分類番号 5 桁までの数字
s24	s21 の自立語の分類語彙表での分類番号 3 桁までの数字
s25	表記選択を行う対象の単語を含む文節に係る文節の最後の自立語
s26	s25 の品詞
s27	s25 の自立語の分類語彙表での分類番号 5 桁までの数字
s28	s25 の自立語の分類語彙表での分類番号 3 桁までの数字
s29	表記選択を行う対象の単語を含む文節に係る文節の自立語
s30	s29 の品詞
s31	s29 の自立語の分類語彙表での分類番号 5 桁までの数字
s32	s29 の自立語の分類語彙表での分類番号 3 桁までの数字
s33	表記選択を行う対象の単語を含む文節に係る文節の最初の付属語
s34	s33 の品詞
s35	表記選択を行う対象の単語を含む文節に係る文節の最後の付属語
s36	s35 の品詞
s37	表記選択を行う対象の単語を含む文節に係る文節に含まれる付属語
s38	s37 の品詞
s39	表記選択を行う対象の単語を含む文節に係る文節に含まれる記号
s40	s39 の品詞
s41	表記選択を行う対象の単語を含む文節が係る文節の最初の自立語
s42	s41 の品詞
s43	s41 の自立語の分類語彙表での分類番号 5 桁までの数字
s44	s41 の自立語の分類語彙表での分類番号 3 桁までの数字
s45	表記選択を行う対象の単語を含む文節が係る文節の最後の自立語
s46	s45 の品詞
s47	s45 の自立語の分類語彙表での分類番号 5 桁までの数字
s48	s45 の自立語の分類語彙表での分類番号 3 桁までの数字
s49	表記選択を行う対象の単語を含む文節が係る文節の自立語
s50	s49 の品詞
s51	s49 の自立語の分類語彙表での分類番号 5 桁までの数字
s52	s49 の自立語の分類語彙表での分類番号 3 桁までの数字
s53	表記選択を行う対象の単語を含む文節が係る文節の最初の付属語
s54	s53 の品詞
s55	表記選択を行う対象の単語を含む文節が係る文節の最後の付属語
s56	s55 の品詞
s57	表記選択を行う対象の単語を含む文節が係る文節の付属語
s58	s57 の品詞
s59	表記選択を行う対象の単語を含む文節が係る文節の記号
s60	s59 の品詞
s61	表記選択を行う対象の単語の前 1,2,3,4,5 文字
s62	表記選択を行う対象の単語の後 1,2,3,4,5 文字

表 2: 機械学習の再現率の高さごとの割合

再現率の高さ	割合
高	8.63 % (81/939)
中	16.40 % (154/939)
低	74.97 % (704/939)

・ ぜひ：是非

(正解例 1a) ウサギは捕まえられず、幼稚園児の末娘に子供用バニーガールの格好をさせたら、１２年後にも ぜひ と好

評だった。

(正解例 1b) 物事の 是非 を知り、善悪の基準について考える機会を尊ぶ世の中になってほしいと願う。

(失敗例 1a) ぜひ また会いたいと話していた。

(失敗例 1b) 婚前交渉の 是非、恋愛結婚と見合い結婚、学校の恋愛文化、自由恋愛、晩婚非婚化など、文学、テレビドラマから昭和の恋愛を検証する。

表 3: 機械学習の結果 (ぜひ：是非)

表記	再現率	適合率	総数
ぜひ	99.17 %	98.28 %	1442
是非	98.48 %	99.26 %	1642

表 4: 機械学習が参考にした素性 (ぜひ：是非)

ぜひ		是非	
素性	正規化 α 値	素性	正規化 α 値
s15：と	0.8834	s61：の	0.9591
s61：を	0.7600	s62：この本	0.7234
s17：と	0.7406	s15：も	0.7008

この「ぜひ：是非」は再現率高の単語である。そして、機械学習が参考にした素性と正解例を見てみると、素性となっている文字が存在している。例えば、(正解例 1a) では、「ぜひと好評」の部分で「s15：と」があり、(正解例 1b) では、「物事の是非」の部分で「s61：の」がある。これらの素性が正規化 α 値が高い素性となったのは、「ぜひ」は、「ぜひ と 勧められて」というような使われ方や、「… を ぜひお願いします」というような使われ方をされ、「是非」は、「物事 の 是非」というような使われ方や、「… の是非 も 問われる」というような使われ方をされることが多いためと思われる。このことと機械学習の再現率が高いことから、「ぜひ：是非」は素性によって適切な表記選択を行える単語であると推測できる。また、「ぜひ：是非」は副詞で用いられているか名詞で用いられているかによって使い分けられていると考えられるので、人間による判定でも容易であると推測できる。

・ 引き揚げる：引き上げる

(正解例 2a) ９７年のアジア金融危機では、短期の投機資金がタイやインドネシアなどから 引き揚げ られ、新興諸国の通貨が暴落する事態になった。

(正解例 2b) 当初、０４年中に日米で各１００万台、計２００万台だった販売計画を、２８０万台に 引き上げ、無事達成できた。

(失敗例 2a) 派遣していた役員２人も 引き揚げる 方針。

(失敗例 2b) 支払限度額は加入者増に応じて 引き上げて おり、現在は５兆円に設定、過去最大の関東大震災級の被害でも対応できるようにしている。

表 5: 機械学習の結果 (引き揚げる：引き上げる)

表記	再現率	適合率	総数
引き揚げる	67.23 %	82.99 %	537
引き上げる	97.20 %	93.59 %	2642

表 6: 機械学習が参考にした素性 (引き揚げる:引き上げる)

引き揚げる		引き上げる	
素性	正規化 α 値	素性	正規化 α 値
s29：資金	0.8199	s62：幅	0.8407
s62：船	0.7963	s32：137	0.8238
s62：者	0.7844	s61：に	0.7525

この「引き揚げる：引き上げる」は再現率中の単語である。データ文を見てみると、「引き揚げる」は (正解例 2a) のように、「資金 (投資) を引き揚げる」といった使い方をされるものが多かった。そのため、「s29：資金」が最も強い素性になったものと考えられる。一方、「引き上げる」は (正解例 2b) のように、「(… の値) に引き上げる」といった使い方がされるものが多かった。そのため、「s61：に」が素性として参考にされていると考えられる。また、「引き揚げる」は「もとの所へ戻す。取り戻す。撤退する。」などの意味で用いられており、「引き上げる」は「程度や値を高くする。」といった意味で用いられているものと考えられる。そのため、人間でも判定は容易であると推測できる。

・ 盛りつける：盛り付ける

(正解例 3a) パックで買った総菜も必ず皿に 盛りつける。

(正解例 3b) 彩りよく 盛り付けて ボリューム感を出す。

(失敗例 3a) 切って混ぜて 盛りつける だけで、本格レストランの味が楽しめた。

(失敗例 3b) 日本の四季にはぐくまれた恵みが食卓に豊かな季節を 盛り付けた。

表 7: 機械学習の結果 (盛りつける：盛り付ける)

表記	再現率	適合率	総数
盛りつける	28.57 %	29.63 %	28
盛り付ける	44.12 %	42.86 %	34

表 8: 機械学習が参考にした素性 (盛りつける:盛り付ける)

盛りつける		盛り付ける	
素性	正規化 α 値	素性	正規化 α 値
s50：名詞	0.6178	s38：助詞	0.5838
s42：名詞	0.6178	s36：助詞	0.5838
s46：名詞	0.6008	s34：助詞	0.5838

この「盛りつける：盛り付ける」は再現率低の単語である。そして、機械学習が参考にした素性を見ても特に目立った素性はなく、再現率も低い。正解例、失敗例を見てみると、「盛り付ける」と「盛りつける」は、「付ける」を漢字で書くか仮名で書くかの違いでしかなく、単語の意味には違いは見つけられない。そのため、機械学習はそれぞれの判定に参考にする素性を見つけることができず、高い再現率で判定することができなかったものと推測できる。また、人間による判定も非常に困難であると推測できる。

4.3.2 再現率の高さごとの傾向

機械学習の再現率の高さごとの分類で再現率高としたものには、前節で示したように、素性によって表記選択を行える単語があった。「是非：ぜひ」の例のように、データ文中に機械学習が判定の参考にした素性があることが確認できた。さらに、それぞれの表記は、文脈や単語の意味ごとに使い分けられていた。これは、機械学習が表記選択に役立つ素性を適切に認識し、その素性を有効に利用することによって正しく表記選択を行えることを意味している。また、これらの単語は、文脈や単語の意味ごとに使い分けがされているので、人間でも容易に判定できる。また、「稲穂：いなほ」という単語において、「いなほ」が特急電車である「いなほ」として使われている場合があった。このように、一方の表記は人名や固有名詞など特定の使われ方をされており、もう一方の表記では普通名詞として使われている場合、再現率高に分類されていた例も多く見られた。

再現率低と分類されていた単語には、その単語の意味自体に違いが見つけられないものが多く見られた。例えば、「切り詰める：切りつめる」という単語は文によって使い分けが発生するような状況は考えにくい。しかし、新聞や論文といった文書に使用する際には、漢字で書くほうが格好の点で良いと考えられる。新聞において漢字で書くほうが良いというのは、使用頻度から見て取れる。「切り詰める：切りつめる」の使用頻度はそれぞれ「73：16」となっており、「切り詰める」が圧倒的に多い。このため、再現率も「切り詰める」が 97.26%、「切りつめる」が 12.50%のように、使用頻度の高いほうが再現率が高い結果になったものと思われる。極端なものでは、片方が 100%でもう片

方が 0%であるものもあった (「取って代わる：とって代わる」「独り暮らし：ひとり暮らし」など)。再現率と使用頻度の関係も同様に、使用頻度が高いほうが再現率が高い結果になっている。全体の傾向として、「切り詰める：切りつめる」のように一部の表記が違うものや、漢語の一部を仮名で書く交ぜ書きをしているもの (「謙遜：謙そん」「洞くつ：洞窟」) が多く見られた。また、外来語を片仮名で表記することによって生じるゆれも、再現率低として分類されているものが多かった。外来語の例では、「ディスプレー：ディスプレイ」「ルネサンス：ルネッサンス」などがある。他にも、動物や魚、植物などの名前の表記の違い (「カメ：亀」「ふぐ：フグ」「キノコ：きのこ」) や、擬音語など (「ふらふら：フラフラ」「ぞっと：ゾッと」) も再現率低として分類されているものが多かった。これらの単語は、表記が違うだけで単語の意味そのものは同じであるという共通点がある。こうした単語は、機械学習による判定で高い再現率で表記選択を行えなかった。また、人間による表記選択も非常に困難であると推測できる。

機械学習の再現率の高さごとの分類で再現率中としたものの中にも、再現率高と同じく、文脈や単語の意味ごとに表記選択ができているものがあった。例えば、前節の例で示した「引き揚げる：引き上げる」がこれにあたる。「引き揚げる：引き上げる」の再現率は、同じ再現率中の中でも高いほうである。再現率中に分類されいていた単語の傾向としては、比較的再現率の高いものには再現率高のものと同様の傾向が見られ、比較的再現率の低いものには再現率低のものと同様の傾向が見られた。

再現率の高さごとの傾向として、以上のような傾向が見られたが、再現率低に分類されることの多かった外来語や動植物の名前も、一部再現率高として分類されているものがあった。これは、再現率高の他の単語と同様に、高い再現率で判定された何らかの理由があったためである。例えば、「エンターテインメント：エンタテインメント」という単語は、「エンタテインメント」が「ソニーコンピュータエンタテインメント」などの企業名として使われており、また、企業名が使用されている文も多かったため、高い再現率で判定されたものと推測できる。

5 おわりに

本研究では、表記のゆれの問題を扱った。表記のゆれがある単語を文中で用いる場合、どの表記を用いるかは判断に迷うことが多い。どのような単語が表記選択が容易で、どのような単語が表記選択が困難か、その傾向や特徴を捉えられれば、今後の表記のゆれの研究やシステムにも役立つと考えた。そこで、本研究では、人間による表記選択が容易な単語とそうでない単語を、機械学習を用いて推定することを行った。機械学習によって高い再現率で表記選択を行えたものは人間による表記選択も容易で、機械学習によって高い再現率で表記選択を行えなかったものは人間による表記選択も困難であるということを確認するため、表記のゆれがある単語について機械学習による表記選択を行った。機械学習法として最大エントロピー法を用いた。

機械学習による判定で、高い再現率で表記選択を行えたものと高い再現率で表記選択を行えなかったものについて、なぜそのような結果になったのかを考察した。その結果、高い再現率で表記選択を行えたものは、機械学習が単語の意味ごとに共起する素性によって表記選択をしていることがわかった。例えば、「ぜひ：是非」という単語では、「ぜひ」では「s15：と」や「s61：を」という素性が正規化 α 値が高く、「是非」では「s61：の」や「s15：も」が正規化 α 値が高いという結果が得られた。これは、「ぜひ」は、「ぜひ と 勧められて」というような使われ方や、「… を ぜひお願いします」というような使われ方をされ、「是非」は、「物事 の 是非」というような使われ方や、「… の是非 も 問われる」というような使われ方をされるためと思われる。また、この場合、人間でも表記選択が容易であることを確認した。高い再現率で表記選択を行えなかったものは、表記が違うだけで単語の意味そのものに違いが見られず、人間でも表記の選択が困難であるものが多かった。これらの結果から、機械が高い再現率で表記選択を行えたものは人間でも正しく判定することが容易で、機械が高い再現率で表記選択を行えなかったものは人間でも正しく判定することが困難であるものがあることを確認した。

また、本研究の機械学習は表記選択自体にも役立つことを確認した。機械学習による表記選択の実験を行った結果、表記のゆれがある単語 939 語中 81 語について、それぞれの単語の出現頻度数上位 2 つの表記とも 80%以上の再現率で適切な表記を選択できた。

参考文献

[1] 岡部 浩司, 河原 大輔, 黒橋 禎夫, “格フレームを用いたかな表記語の曖昧性解消”, 言語処理学会第 12 回年次大会, pp.1115-1118, (2006).

[2] 岡部 浩司, 河原 大輔, 黒橋 禎夫, “代表表記による自然言語リソースの整備”, 言語処理学会第 13 回年次大会, pp.606-609, (2007).

[3] 黒橋 禎夫, 河原 大輔, “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).

[4] 黒橋 禎夫, 河原 大輔, “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).

[5] 西川 彩, 西村 涼, 渡辺 靖彦, 岡田 至弘, “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, pp.729-732, (2009).

[6] 服部 弘幸, 関 和弘, 上原 邦昭, “英音素変換を用いたカタカナ異表記の自動生成”, 情報処理学会研究報告 自然言語処理研究会報告 2007(94), pp.59-64, (2007).

[7] 日本 満, 明関 幸江, “同音類義語「かえる」の漢字 (変・替・換・代) 表記のゆれについての一考察：新聞コーパスとアンケート調査を基に” 名古屋市立大学人文社会学部研究紀要第 17 号, pp.187-200, (2004).

[8] 村田 真樹, 内元 清貴, 馬 青, 井佐原 均, “機械学習手法を用いた名詞句の指示性の推定”, 自然言語処理 (言語処理学会誌) 7 巻 1 号, pp.31-50, (2000).

[9] 村田 真樹, 神崎 享子, 内元 清貴, 馬 青, 井佐原 均, “意味ソート msort”, 情報処理学会研究報告 自然言語処理研究会報告 99(22), pp.89-96, (1999).

[10] 村田 真樹, 西村 涼, 金丸 敏幸, 土井 晃一, 鳥澤 健太郎, “ユーザ個人の興味の影響を考慮した情報の重要度を決める要因の抽出・分析”, 言語処理学会第 15 回年次大会, pp.554-557, (2009).

付録

本研究で使用した表記のゆれがある単語 939 個のうち一部を、機械学習の再現率の高さごとにそれぞれ表 9、表 10、表 11 に示す。

表 9: 再現率高

表記	表記	表記
不和だ：ふわだ	稲穂：いなほ	高々：たかだか
瑠璃：るり	そうそう：草々	かわれる：変われる
こも：薦	手引き：手引	組み手：組手
家々：いえいえ	そぶり：素振り	深層：深そう
噴水：ふんすい	清い：きよい	うかる：受かる

表 10: 再現率中

表記	表記	表記
冬物：冬もの	小道：こみち	利己：りこ
讃歌：賛歌	不意だ：ふいだ	うっぷん：うっ憤
きび：機微	ドブ：どぶ	鶯：ウグイス
裸足：はだし	肩透かし：肩すかし	メジロ：目白
のら：野良	ハッと：はっと	いばら：茨

表 11: 再現率低

表記	表記	表記
逃げ惑う：逃げまどう	すそ：裾	物まね：ものまね
姉ちゃん：ねえちゃん	朝顔：あさがお	けん怠：倦怠
刷り込む：すり込む	うす：臼	轍：わだち
投函：投かん	内証：内緒	すれる：擦れる
そむける：背ける	研ぐ：とぐ	サメ：鮫