複数分野の文書を用いた日本語誤り表現の検出

田中 駿 † 村田 真樹 †† 徳久 雅人 ††
† 鳥取大学工学部知能情報工学科
†† 鳥取大学大学院工学研究科情報エレクトロニクス専攻
{s082031,murata,tokuhisa}@ike.tottori-u.ac.jp

1 はじめに

近年，パソコンやインターネットの普及により，計算機を使って文字を入力する機会が増えている．また，ブログ等の気軽に文書を書ける媒体が出現したことによって，口語的表現や，くだけた表現，誤った表現等をよく目にする．そういった表現は研究やビジネス等に用いられる正式な文書作成時には不適切であるため，それらを検出することが望まれる．

先行研究として，既に入力誤り検出・表記統一を目的とした研究が行われている [1, 2, 3, 4, 5, 6, 7]．例えば，白木ら [8] は平仮名列を抽出し，辞書データベースと照合することでスペルチェックを行っている．

しかしこれらの手法では誤り (誤字や脱字等) を検出できても，くだけた表現や，一般にその文書では利用されることが少ない (好ましくない) 表現を検出することが難しいという問題がある．そこで本研究では，複数分野の文書を用いて当該分野において不適切となる表現の検出を行う．本稿では，同種の文体や表現を利用する文書群が属するものを「分野」と呼ぶことにしている．

本研究では，白木ら [8] と異なり，平仮名列ではなく，付属語と接続詞と感動詞と連体詞と句読点の連続 (以下，これらを対象語列とする) を抽出し，複数分野の文書での対象語列の出現頻度を利用して誤り表現の検出を行う．

本研究の主な点をあらかじめ整理すると以下のようになる．

- 日本語誤り検出に対して，複数分野の文書での出現頻度を利用するという特徴的な手法を提案している．
- 提案手法は，具体的には，修正する文書と同じ分野の文書での頻度が小さく，修正する文書と異なる分野の文書での頻度が大きい表現を誤りとするものである．簡単に言えば，例えば新聞において新聞で頻度が少なくブログで頻度の高い表現があった場合にブログにあるようなくだけた表現の可能性が高いとして誤りとするものである．
- 実験において，複数分野の文書での出現頻度を利用する方 (提案手法) が利用しないよりも，統計的検定により有意に少ない誤検出で (高い適合率で) 誤り検出をできることを確認した．
- 提案手法は残念ながら，誤り表現の検出の再現率，適合率，F 値は低かった．複数分野の文書での出現頻度を利用すると良いことまではわかったが，それを有効利用するには至っていない．

表 1 新聞での出現が 1 回の対象語列の出力例

fr_b	対象語列	対象語列の品詞情報
0	をめぐって +、+ また	助詞-格助詞-連語 + 記号-読点 + 接続詞
0	など + は +、+ そうした	助詞-副助詞 + 助詞-係助詞 + 記号-読点 + 連体詞
0	で +、+ あるいは +、	助動詞 + 記号-読点 + 接続詞 + 記号-読点
1	だけ + で + あろ + う +。	助詞-副助詞 + 助動詞 + 助動詞 + 助動詞 + 記号-句点
1	べき + で + あっ + て +、	助動詞 + 助動詞 + 助動詞 + 助詞-接続助詞 + 記号-読点
2	にあたって +、+ この	助詞-格助詞-連語 + 記号-読点 + 連体詞
2	た + って +、+ この	助動詞 + 助詞-格助詞-連語 + 記号-読点 + 連体詞
98	ごめんなさい +。	感動詞 + 記号-句点
148	っていう + か	助詞-格助詞-連語 + 助詞-副助詞／並立助詞／終助詞
249	です + か + ね +。	助動詞 + 助詞-副助詞／並立助詞／終助詞 + 助詞-終助詞 + 記号-句点
264	やん	助動詞
360	か + なぁ +。	助詞-副助詞／並立助詞／終助詞 + 助詞-終助詞 + 記号-句点
948	だ + よ + ね	助動詞 + 助詞-終助詞 + 助詞-終助詞
1265	まぁ +、	感動詞 + 記号-読点

2 提案手法

前述の通り，提案手法の特徴は複数分野の文書での頻度を利用することである．修正対象の文書と同一の分野の文書での出現頻度が低く，かつ，修正対象と異なる分野の文書での出現頻度が高い対象語列[*1]が出現するとそれを誤りと判定する．

検出は以下の手順で行う．

手順 1 ChaSen[9] による形態素解析で，分野の異なる 2 つの文書 (ここでは文書 A，文書 B とおく．

[*1] 本稿では対象語列を付属語列等，つまり内容語以外としている．これは，内容語を対象とすると，文体や表現による誤りだけでなく，文章の内容の違いだけでも誤りとしてしまうためである．しかし，内容語であっても該当分野で利用すると不適切な表現もあり，内容語も利用していくという手法の拡張も考えられる．

できるだけ大量のデータであることが望ましい) から対象語列を抽出する．

手順２　それぞれの文書での対象語列の出現頻度を調べる (例として表 1 を参照．fr_b はブログの文書での頻度)．

手順３　文書 A と同じ分野の文書 X で誤り検出をしたい場合，文書 A での出現が 0 回であり，文書 B での出現が多い対象語列を文書 X で探し，見つかればそれを誤りの可能性があるものとして出力する．

3　実データを用いた実験

3.1　実験方法

2 節の提案手法の性能を次のように調べる．文書 A として新聞，文書 B としてブログ (または文書 A をブログとして，文書 B を新聞とする．また，文書 A と同じ分野の文書として文書 X を設定する．) を使用し，2 節の手順 1 から 3 を行う．ただし文書 A での頻度を 2 節の手順 3 では 0 回としているが，ここでは 1 回として抽出する[*2]．ここでは頻度情報を取り出すデータには訓練データだけでなく評価用データ 1 個も含まれているので，新聞の頻度は 0 でなく 1 を使用する．人手による評価の方法を以下に示す．

- 通常の文で，正しく使用されているものは○と判定する．
- 鍵括弧は利用していないが，引用などにより意図的に文体を変えている箇所は△と判定する．
- 明らかな誤りは×と判定する．

用いる新聞とブログの文書は毎日新聞 1991 年 (1 年分,7171 記事) と，ブログサイト「ココログ」の 2009 年 11 月 1 日〜7 日に書かれた記事から新聞と同量抜き出したものである．

3.2　実験結果

対象語列は，新聞からは 15,926 種類，ブログからは 48,531 種類抽出でき，重複を除いてあわせると，全体では 55,958 種類が得られた．3.1 節に基づいて処理した結果を表 2 に示す (ただし表中の fr_b はブログでの出現頻度である)．表 2 は $fr_b = 0$ のデータからランダムに 100 個，$fr_b \geqq 1$ のデータからランダムに 100 個の対象後列をを抜き出して，新聞での使われ方を評価した結果である．

新聞において△と判定された対象語は，記事中では鍵括弧は使用しないが，引用などに出現した文であり，実際には誤りの文ではない．しかし，△と判定された対象語はブログなどで使用されるくだけた日本語であり，その検出個数を調べることで，そのような表現の抽出性能を調べることができる．このため，ここではくだけた日本語である△と，誤り表現である×を検出できると，検出成功と考えて評価した．

表 2 より，△または×の検出の割合 (適合率) は文書 B での頻度 (fr_b) が増えるに従い上昇することが確認された．これにより，提案手法の有効性が確かめられた．

表 2 の結果については表 3 を利用して χ^2 検定を行って，ブログの頻度が $fr_b \geqq 2$ の場合と $fr_b \leqq 1$ の場合とで，△ or ×の検出の割合 (適合率) に有意差があることを確認した．すなわち，ブログ (文書 B) の頻度を利用する提案手法の有効性は統計的検定によっても確認された．ここで表 3 は，fr_b を 1 以下と 2 以上に整理し，△と×の評価を 1 つにまとめたものである．

同様に，文書 A と X にブログを利用し文書 B に新聞を利用して評価と検定を行うと，表 4 と表 5 の結果が得られた．この結果でも，新聞 (文書 B) の頻度が上昇するほど，△ or ×の検出の割合 (適合率) が上昇している．χ^2 検定により，新聞 (文書 B) での頻度の 1 以下と 2 以上と，評価○と△ or ×とは独立でない (関係がある) ことがわかった．よって，この実験でも提案手法が有意に有効であることが確認された．

表 2　新聞での出現が 1 回の対象語列の評価結果

評価	$fr_b = 0$	$fr_b = 1$	$fr_b = 2$	$fr_b \geqq 3$
○	79%(79/100)	79%(30/38)	64% (7/11)	41%(21/51)
△	21%(21/100)	21%(8/38)	27%(3/11)	59%(30/51)
×	0%(0/100)	0%(0/38)	9%(1/11)	0%(0/51)

表 3　表 2 を 2*2 分割表に変換したもの

評価	$fr_b \leqq 1$	$fr_b \geqq 2$	行計
○	109	28	137
△ or ×	29	34	63
列計	138	62	200

表 4　ブログでの出現が 1 回の対象語列の評価結果

評価	$fr_b = 0$	$fr_n = 1$	$fr_n = 2$	$fr_n \geqq 3$
○	88%(88/100)	84%(51/61)	73%(11/15)	50%(12/24)
△	11%(11/100)	15%(9/61)	27%(4/15)	50%(12/24)
×	1%(1/100)	1%(1/61)	0%(0/15)	0%(0/24)

表 5　表 4 を 2*2 分割表に変換したもの

評価	$fr_n \leqq 1$	$fr_n \geqq 2$	行計
○	139	23	162
△ or ×	22	16	38
列計	161	39	200

4　擬似的に作成したデータを用いた実験

4.1　入れ替えた文の検出

別の新聞記事とブログ記事として，毎日新聞 1992 年とココログの 2009 年 10 月の記事からそれぞれ 10,000 文を用意し，ランダムに 1,000 文を入れ替え，入れ替えた文をどのくらい正しく 2 節の提案手法で検出できるかを調べる．提案手法の手順は 2 節の通りであるが，

[*2] これは Leave one out 法という，「N 個のデータについて考える場合に，それを $N-1$ 個の訓練データと 1 個の評価用データとに分割し $N-1$ 個の訓練データを用いた学習結果で 1 個の評価用データを評価する」という概念に基づいているからである．

表 6　新聞頻度 0 で検出した結果をブログ頻度で分けたもの

	$fr_b \leqq 1$	$fr_b \geqq 2$	行計
混ぜた文	3	3	6
もとの文	314	8	322
列計	317	11	328

表 7　ブログ頻度 0 で検出した結果を新聞頻度で分けたもの

	$fr_n \leqq 1$	$fr_n \geqq 2$	行計
混ぜた文	5	2	7
もとの文	320	8	328
列計	325	10	335

3 節とは違い，検出を行うデータが頻度を算出するデータと異なっている点に注意する (2 節のとおり検出には文書 A の頻度 0 を利用する).

4.1.1　新聞に 1,000 文のブログ文書を混ぜた実験

文書 A,B に 3 節で用いた新聞とブログの文書を利用した．文書 X には，新聞の文書に 1,000 文のブログの文書を混ぜたデータを利用した (ただし文書 A,B と文書 X に重なりはない．これは 4.1.2 節でも同様である．)．新聞での頻度が 0 の対象語列を抽出すると，混ぜた 1,000 文からは 6 文が，もとの 9,000 文からは 322 文が検出された．これら 328 文を，ブログでの頻度，および，混ぜた文かいなかで分けると表 6 のようになる．3 節と同様に χ^2 検定により，ブログ (文書 B) での頻度が 2 以上の方が，1 以下のものよりも，有意に混ぜた文の検出の割合 (適合率) が高いことが確認された．よって，ブログ (文書 B) での頻度を利用することの有効性が確認された．

4.1.2　ブログに 1,000 文の新聞の文書を混ぜた実験

文書 A,B に 3 節で用いたブログと新聞の文書を利用した．文書 X には，ブログの文書に 1,000 文の新聞の文書を混ぜたデータを利用した．ブログでの頻度が 0 の対象語列を抽出すると，混ぜた 1,000 文からは 7 文が，もとの 9,000 文からは 328 文が検出された．これら 335 文を，新聞での頻度と混ぜた文かいなかで分けると表 7 のようになる．同様に χ^2 検定を行うことにより，新聞 (文書 B) での頻度を利用することが有意に有効であることが確認された．

4.2　考察

4.1.1 節と 4.1.2 節の結果より，両方の結果で文書 B での頻度を用いる提案手法が有効であることがわかった．例として，ブログに新聞を混ぜた実験 (4.1.2 節) で提案手法 (ここでは $fr_b = 0$ かつ $fr_n \geqq 2$ とする) により正しく混ぜた文を検出できた対象語列を表 8 に示す．これは表 7 の $fr_n \geqq 0$ であり混ぜた文である 2 個に相当する．この例をみると，ブログに似つかわしくない堅めの表現が正しく取り出せていることがわかる．同じ実験で提案手法 ($fr_b = 0$ かつ $fr_n \geqq 2$) により正しくもとの文を誤りとはしなかった (もとの文を取り出さなかった) 場合の対象語列を表 9 に示す．これらの例は，新聞での頻度が少なく，誤り表現としては取り出されなかった．これらの例はブログにあってもおかしくない表現であり，提案手法は正しく誤り表現としていないことがわかる．

表 8　4.1.2 節で提案手法で正しく検出できた対象語列 ($fr_b = 0$ かつ $fr_n \geqq 2$ のもの)

fr_n	対象語列	対象語列の品詞情報
6	だ＋が＋、＋さて	助動詞＋助詞-接続助詞＋記号-読点＋接続詞
4	ない＋か＋、＋と＋の	助動詞＋助詞-副助詞／並立助詞／終助詞＋記号-読点＋助詞-格助詞-引用＋助詞-連体化

表 9　4.1.2 節で提案手法で正しく誤り表現としなかった対象語列 ($fr_b = 0$ かつ $fr_n \leqq 1$ のもの)

fr_n	対象語列	対象語列の品詞情報
1	とか＋で＋は＋なく	助詞-並立助詞＋助動詞＋助詞-係助詞＋助動詞
0	ばっかり＋だっ＋たら	助詞-副助詞＋助動詞＋助動詞

表 10　4.1 節の各実験における再現率・適合率・F 値

節	提案手法			ベースライン		
	再現率	適合率	F 値	再現率	適合率	F 値
4.1.1	0.003	0.273	0.006	0.006	0.018	0.016
4.1.2	0.002	0.200	0.004	0.007	0.021	0.011

次に 4.1.1 節と 4.1.2 節の実験結果について再現率，適合率，F 値を調べた．その結果を表 10 に示す．表 10 中のベースラインは文書 B がどのような頻度であっても誤りとして検出する手法であり，ここでの提案手法は文書 B で頻度 2 以上であったもののみを誤りとして検出するものである．提案手法は，再現率，F 値ではベースラインに劣っている[*3]．適合率は，提案手法はベースラインよりも高いが，値自体は低いものであった．提案手法のように文書 B を考慮することが誤り検出 (誤り検出における適合率の上昇) に有効であることは統計的検定で確認されているが，再現率，適合率，F 値の低さを考えると，提案手法はまだまだ改善の必要性がある．

再現率が特に低かったため，4.1.1 節を例に，実験結果で検出できなかったブログ記事をランダムに 20 件取りだし，どのような文体がどれくらいの割合で含まれているかを調査した．表 11 にその割合を示す．ブログ記事の文は一般に口語的な文であることが想定されるが，『新聞に近い文体で書かれた文』『短い文，助詞を含まない文，名詞のみの文など』が 6 割も含まれていることがわかった．この 6 割のものは検

[*3] 提案手法は，再現率，F 値ではベースラインに劣っているが，適合率で上回っているのでベースラインよりも効果的な利用方法がある．例えば，文書 B での頻度が高い誤りの可能性が高い表現から修正候補として提示し，文書 B での頻度 0 のまですべて表示するとすれば，再現率はベースラインと一致し，かつ，誤りの可能性が高い表現から修正できる．

表 11 検出できなかったブログ記事 20 件に含まれる文体の割合

文体の種類	例	割合
新聞に近い文体で書かれた文	このｂｌｏｇを書きながら太平洋戦争について、今までも色々と記事を書いてきた。	5/20(25%)
短い文，助詞を含まない文，名詞のみの文など	超楽しい！	7/20(35%)
一般的な口語的文	今度こそ、一緒に鍋食べましょうね！	8/20(40%)

出できなくても仕方がないものと見ることができる．これらの 6 割のものを再現率の計算に含めないものとして再計算を行うと，表 10 の提案手法の再現率は $3/((1000-3)*0.4+3))=0.007$ となる．この再計算をしても再現率が低いことに変わりがなかった．

提案手法の結果を改善する方法として次の方法が考えられる．

方法 1 頻度情報を取得するためのデータを，もっと増やす．

方法 2 完全に同じ文体 (論調) のみで構成されているデータを使用する．

方法 1 により，頻度情報を集めたデータに存在しないデータの出力 (文書 A の頻度が 0，文書 B の頻度が 0 のもの) を減らすことができる．現状では，文書 A の頻度が 0，文書 B の頻度が 0 のものの中にも誤りとして検出したいものが数多く含まれている．方法 1 により，それらを検出できるようになる可能性が出てくる．

方法 2 について議論する．現在の実験では，新聞とブログを実験に用いている．厳密には新聞やブログは，様々な文体 (堅めの文章とくだけた文章) が混ざっている[*4]．今回の実験では全紙面，全記事を使用したために，堅めの文章とくだけた文章が混在した状態で頻度を算出していると予想される．方法 2 のように，同じ文体の文書のデータを，データ A やデータ B として利用して実験を行うと性能が上昇すると期待される．

ここでは方法 1,2 を示した．しかし，これだけではまだ性能の高い誤り検出は困難かもしれない．性能をあげるための他の方法も考えていきたい．

5 おわりに

本研究では，複数分野の文書を用いた日本語誤り表現の検出の手法を提案した．提案手法は，具体的には，修正する文書と同じ分野の文書での頻度が小さく，修正する文書と異なる分野の文書での頻度が大きい表現が存在するとそれを誤りとする手法である．

新聞とブログを用いて，実データを用いた実験と疑似データを用いた実験の二種類を行った．この二種類の実験ともに，複数の分野の文書での頻度を利用した方が単一の分野の文書での頻度しか用いない方法よりも，統計的検定により有意に性能 (適合率) が高いことを確認した．これにより，今後は，日本語誤り表現の検出に，複数の分野の文書での頻度を用いていくと良いことがわかった．

しかし，提案手法は，実験の結果，再現率，適合率，F 値が低いことがわかった．複数の分野の文書での頻度を利用した方が良いことはわかったが，それをうまく利用して性能高く誤り検出を行うには至っていない．今後は，性能の高い誤り検出をするのに，複数の分野の文書での頻度をどのように利用すればよいかを検討していきたい．

[*4] 例えば新聞は紙面により文体が異なっており，政治，経済，国際面では堅い表現が使われ，コラム，広告，投書，家庭面等ではくだけた (口語的) 表現がよく使われる．また，新聞中の引用箇所においてくだけた表現が使われる場合もある．ブログについても，堅めの文章とくだけた文章が混在する．ブログの多くは，日常の出来事をまとめたメモや日記であり，文体を気にしていない，くだけた文章で構成されている．その一方で，ニュースや事件についての転載や，自身の意見や感想をまとめた箇所は堅めの記事となっている．

参考文献

[1] 池原悟, 小原永, 高木伸一郎: 文書支援システムにおける自然言語処理, 情報処理, 34(10), pp.1249-1258, 1993.

[2] Masaki Murata, Hitoshi Isahara: Automatic detection of mis-spelled Japanese expressions using a new method for automatic extraction of negative examples based on positive examples, IEICE Transactions on Information and Systems, E85-D(9), pp.1416-1424, 2002.

[3] 村田真樹, 井佐原均: 言い換えの統一的モデル，言語処理学会誌, 11(5), pp.113-133, 2004.

[4] 荒木哲郎，池原悟，塚原信幸：m 重マルコフモデルによる日本語の誤字、脱落及び挿入誤りの検出法, 全国大会講演論文集 第 47 回平成 5 年後期, (2), pp.109-110, 1993.

[5] 三浦雅則, 横山晶一: 留学生の日本語助詞修正システム, 情報処理学会第 71 回全国大会講演論文集, pp.2_279-2_280, 2009.

[6] 今枝恒治, 河合敦夫, 永田亮, 桝井文人: 日本語学習者の作文における格助詞の誤り検出と訂正, 情報処理学会研究報告, 2003-CE(68), pp.39-46, 2003.

[7] 南保亮太, 乙武北斗, 荒木健治: 文節内の特徴を用いた日本語助詞誤りの自動検出・校正, 情報処理学会研究報告. 自然言語処理研究会報告, 2007-NL-181, pp.107-112, 2007.

[8] 白木伸征, 黒橋禎夫, 長尾眞: 大量の平仮名列登録による日本語スペルチェッカの作成, 言語処理学会第 3 回年次大会発表論文集, pp.445-448, 1997.

[9] ChaSen http://chasen.naist.jp/hiki/ChaSen/