概要

インターネットの普及によって，以前に比べ，携帯電話やパソコンで文字を入力する機会が増えている．また，最近ではブログやSNS，掲示板等の，誰でも気軽に利用できる媒体が数多く出現し，そこでは文法や文体を気にすることなく文字が入力され，情報がやりとりされている．そういった表現を日常的に目にすることで，ビジネスや論文等，正式な文書を作成する際に，誤った (あるいは好ましくない) 表現を使ってしまう可能性がある．

本研究では，そういった表現を検出することを目的としている．検出対象としたい文書と同じ分野の文書と，異なる分野の文書の2個をあらかじめ用意しておき，語句の出現頻度を調べて，検出対象とする分野ではほとんど使われないが，違う分野では多く使われる表現があれば，それは誤りである可能性が高いと判断をする．

本研究では，異なる分野での頻度を利用することが有効であるかを，実データと，擬似的に作成したデータの2つで実験を行い明らかにしている．また，手法の改善点や，利用法についても考察を行っている．

本研究では，修正対象の文書を，新聞の経済面のみとした場合の疑似データに基づく実験も行った．この実験では新聞データにブログの文を混ぜておき，新聞データからブログデータを検出できればブログのようなくだけた文を検出できたという意味で検出に成功したと考える．その実験では，複数の文書の頻度を用いないベースラインの方法 (ブログでの頻度を用いない方法) において，ブログ文の検出は0.299の適合率であった．また，複数の文書の頻度を用いる提案手法 (ブログでの頻度を用いる方法) において，ブログでの頻度が1以上，2以上，10以上とした場合にそれぞれ0.344，0.377，0.468という適合率を得た．提案手法はベースラインよりも高い適合率であった．提案手法はブログでの頻度が高い場合ほど高い適合率を得ることを確認した．

ブログでの頻度が高い場合ほど高い適合率を得ることができるので，提案手法は効果的な利用方法が考えられる．例えば，ブログでの頻度が高い表現から修正候補として提示することで，誤りの可能性が高い表現から人手でチェックしていくことが可能である．

目 次

表 目 次

図 目 次

第1章　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．

本研究の主張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．

- 本研究では，修正対象の文書を新聞の経済面のみとした場合の疑似データに基づく実験も行った．この実験では新聞データにブログの文を混ぜておき，新聞データからブログデータを検出できればブログのようなくだけた文を検出できたという意味で検出に成功したと考える．実験結果は，複数の文書の頻度を用いないベースラインの方法 (ブログでの頻度を用いない方法) において，ブログ文の検出は 0.299 の適合率であった．一方，複数の文書の頻度を用いる提案手法 (ブログでの頻度を用いる方法) において，ブログでの頻度が 1 以上，2 以上，10 以上とした場合にそれぞれ 0.344，0.377，0.468 という，ベースラインよりも高い適合率が得られた．よって提案手法は，ブログでの頻度が高い場合ほど高い適合率を得ることを確認した．

- ブログでの頻度が高い場合ほど高い適合率を得ることができるので，提案手法は効果的な利用方法が考えられる．例えば，ブログでの頻度が高い表現から修正候補として提示することで，誤りの可能性が高い表現から人手でチェックしていくことが可能である．

第2章　関連研究と本研究の関係

2.1　関連研究

この節では日本語誤り検出，統一 (スペルチェック) に関する研究を説明する．

1 章でも述べたが，白木ら [8] は，日本語の中でも平仮名の誤字・脱字に対するスペルチェッカを作成している．平仮名を対象としているのは，漢字の場合は，かな漢字変換が一種の確認作業となるため誤りが少なく，一方平仮名は，かな漢字変換では処理されず，誤字・脱字などの単純な入力ミスが多く残る傾向があるためである．

スペルチェックの方法として，まず大量の辞書構築用テキストから平仮名列を抜きだして，ハッシュ等を使用して辞書データベースを作成する．次にスペルチェック対象テキストから平仮名列のみ抜きだし，辞書データベース中に平仮名列があるかどうかを調べる．なければ誤りの候補としてユーザに示す，という手法である．あらかじめ大量のテキストから抜き出した情報をもとに誤りを探すという点で，本研究と近いと言える．

この他にも日本語誤りを検出する研究は多数あり，高い性能のシステムが提案されている．以下に，本研究を進める際に参考とした研究をまとめる．

池原ら [1] の文献により，誤りの種類とそれに対する検出技術，実用化されたシステムが紹介されている．

村田ら [2, 3] は，正例 (positive examples) から負例 (negative examples) の予測方法を提案している．正例から判定する事例の一般的確率を求め，その確率が高いが正例に出現しないものを負例としている．また，日本語誤り検出と外の関係の文の自動抽出を行っており，有効性と汎用性を示している．

近年の研究では，留学生等の日本語学習者を対象とする研究があり，三浦ら [5] の格フレームを用いた助詞修正システムや，今枝ら [6] の格助詞を対象とした誤り検出と訂正の研究がある．

2.2　関連研究と本研究の違い

本研究は，複数の分野の文書を用いて頻度情報を収集している点が特徴である．また，対象語列 (先述のとおり，付属語と接続詞と感動詞と連体詞と句読点の連続) を検出対象としている．

1 つの文書内の頻度を用いた研究は行われているが，複数の文書の頻度を用いている研究は，調査した限りでは確認していない．よって本研究は，まず複数の文書の頻度が本当に検出に有効かを明らかにし，その上で性能を求めることとする．

第3章　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表現が出現するとそれを誤りと判定する．

3.1 節では検出の対象とする品詞について，3.2 節では提案手法の手順について述べる．

3.1　検出対象の品詞

本研究では付属語 (助詞・助動詞)，接続詞，感動詞，連体詞，句読点の連続している部分を検出の対象とし，以降，これらを 対象語列 と呼ぶ．

対象語列は付属語列等，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 (例えば専門性の高い動詞や名詞を含む場合が考えられる)．

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．

3.2　手順

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．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 3.1 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．

ここで，対象語列の取り出し方について，図 3.1 の例文を用いて説明する．図 3.1 の例文に対して，ChaSen によって形態素解析を行ったときの出力は，図 3.2 の通りである．

この 学校 の 敷地 を 突っ切って い た の かも しれ ませんね。

図 3.1　例文

この コノ この 連体詞
学校 ガッコウ 学校 名詞-一般
の ノ の 助詞-連体化
敷地 シキチ 敷地 名詞-一般
を ヲ を 助詞-格助詞-一般
突っ切っ ツッキッ 突っ切る 動詞-自立 五段・ラ行 連用タ接続
て テ て 助詞-接続助詞
い イ いる 動詞-非自立 一段 連用形
た タ た 助動詞 特殊・タ 基本形
の ノ の 名詞-非自立-一般
かも カモ かも 助詞-副助詞
しれ シレ しれる 動詞-自立 一段 連用形
ませ マセ ます 助動詞 特殊・マス 未然形
ん ン ん 助動詞 不変化型 基本形
ね ネ ね 助詞-終助詞
。 。 。 記号-句点

図 3.2　ChaSen による形態素解析の出力

この例では下線で示した部分が対象語列である．連続しているものについては間に「+」を入れて，文書 A，B での頻度情報と合わせてデータベース化し，検出に用いる．表 3.1 では，4 章で使用している，新聞頻度 1 の対象語列を示している．文書 B の頻度 (ここで

はブログの頻度である fr_b) が高くなるほど，口語的なものが多く含まれていることがわかる．

全体的な流れを，例を用いて図 3.3 に示す．ただし図中での対象語列の頻度は，わかりやすく示すための作例であり，実際の頻度とは異なる．図 3.3 では修正対象を新聞としている．「だよねぇ。」が新聞では出現頻度が 0 回，ブログでは出現頻度が 40 回なので，誤りとして出力をしている．

表 3.1 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
3	ね+ば+、+と	助動詞+助詞-接続助詞+記号-読点+助詞-格助詞-引用
3	へ+の+さらなる	助詞-格助詞-一般+助詞-連体化+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

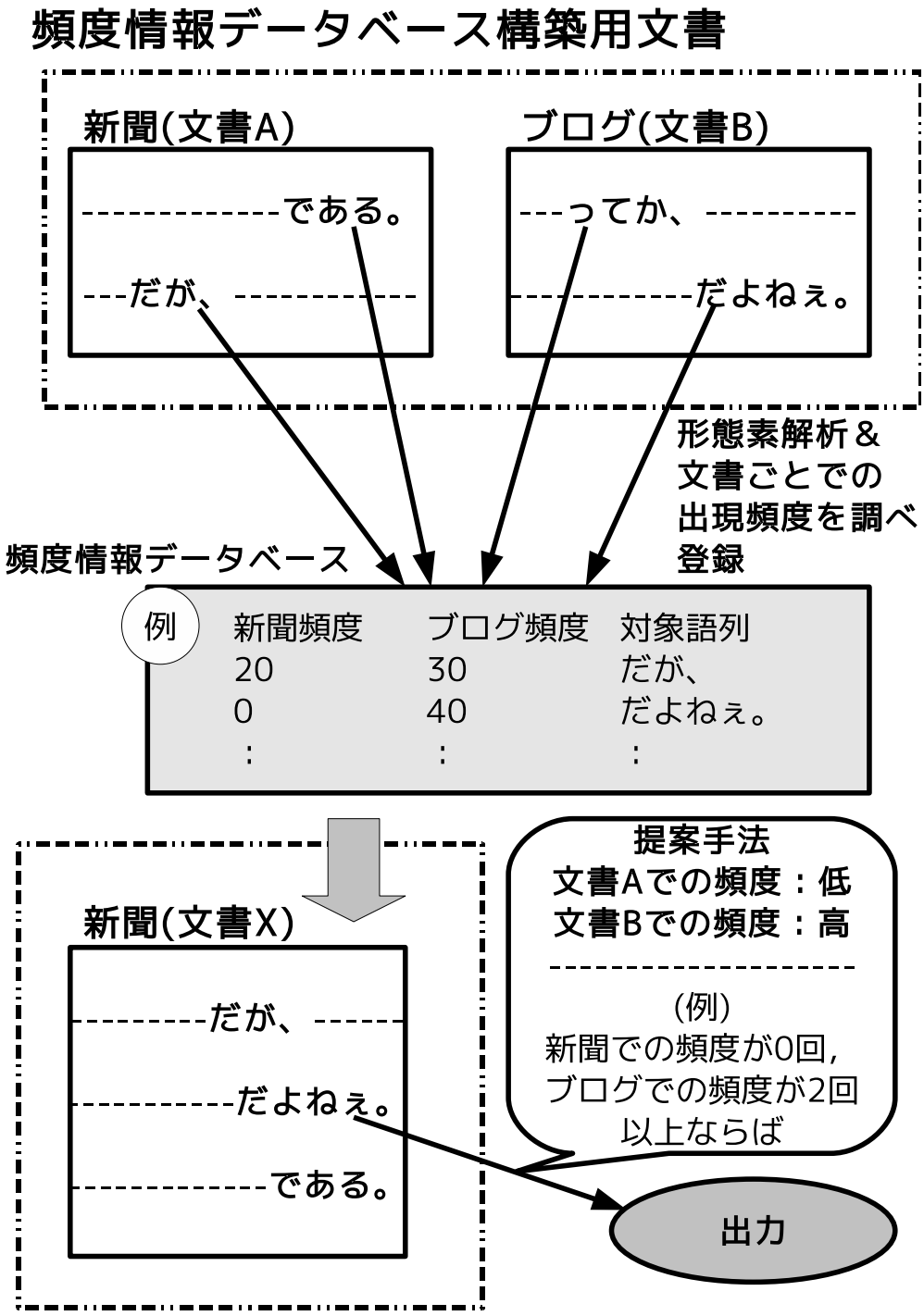
頻度情報データベース構築用文書
新聞(文書A)
ブログ(文書B)
---------である。
---だが、---------
---ってか、---------
---------だよねぇ。
形態素解析＆
文書ごとでの
出現頻度を調べ
登録
頻度情報データベース
例
新聞頻度
ブログ頻度
対象語列
20
30
だが、
0
40
だよねぇ。
：
：
：
新聞(文書X)
---------だが、------
----------だよねぇ。
----------である。
提案手法
文書Aでの頻度：低
文書Bでの頻度：高
(例)
新聞での頻度が0回,
ブログでの頻度が2回
以上ならば
出力

図 3.3 実験の流れ

第4章　実データを用いた実験

4.1　実験方法

3章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，3章の手順1から3を行う．ただし文書Aでの頻度を3章の手順3では0回としているが，ここでは1回として抽出する（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）．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．

4.2 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．4.1 節に基づいて処理した結果を表 4.1 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 4.1 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 4.1 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 4.1 の結果については表 4.2 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．

χ^2 検定について説明する．表 4.1 のように，2*2 の表において，1 行目を左から a,b,g とし，同様に 2 行目を c,d,h，3 行目を e,f,n とすると，検定統計量 T は次のように定義されている．

$$T = \frac{n * (a * d - b * c)^2}{e * f * g * h} \tag{4.1}$$

ここで表 4.2 について，式 (4.1) より，$T = 22.68$ となる．有意水準 $\alpha = 5\%(0.05)$ で自由度 $f = 1$ のとき，棄却域は，χ^2(f,α)=χ^2(1,0.05)=3.84 である (以降の検定も同じ条件で行う)．つまり $T = 22.68 > 3.84$ より，評価○と△ or ×とは独立でない (関係がある) とわかった．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とになる．

表 4.1 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 4.2 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

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を行うと，表 4.3 と表 4.4 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，$T = 25.59 > 3.84$ となったつまり評価○と△ or ×とは独立でない (関係がある) ことになる．よって新聞頻度情報の利用も有効であることがわかった．

これら 2 つの結果より，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 4.3 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 4.4 表 4.3 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

次に，人手評価により×とした事例について考える．×とした事例は表 4.1 と表 4.3 から 3 件見つかった．×とした事例を図 4.1 に示す．図で例文中の該当箇所を [ ] で囲って示す．

表 4.1 で評価×とした『ぬ+で』について，調べた限りでは，このような表現の使われ方はされないことがわかった．

また，表 4.3 では評価×が 2 件見つかった．『を+な+、』は，おそらく尻取りを文字化した文で使用されているのだが，この部分のみ句点が入っており，それにより形態素解析の結果がおかしくなったものと考えられる．『に+で+ある』は周りの文脈から，本来は「～どこ にでもある 小説になってしまう」とするのが正解と判断できる．しかしここでは脱字になっていたため，評価×と判断した．

よって本手法では不適切な文体の表現だけでなく，誤字脱字についても場合によっては検出できることがわかった．

■表 4.1 の評価×
検出した対象語列：ぬ+で
新聞頻度：1 回，ブログ頻度：2 回
品詞：助動詞+助動詞
例文：
毎日新聞社賞- - 股関節柔軟度測定器
左右の股（こ）関節の柔軟度、脚の長さを測定する器具。あお向けに寝て脚を伸ばし、足のかかとをつけ、そのときの左右のつま先の開きの [ぬで] 違いから、股関節の柔軟度を調べられる。

■表 4.3 の評価×
検出した対象語列：を+な+、
新聞頻度：0 回，ブログ頻度：1 回
品詞情報：助詞-格助詞-一般+感動詞+記号-読点
例文：
に：人ぬ：ぬいぐるみね：寝坊の：のるなよは：早すぎひ：引かんようにふ：二人ではへ：平凡にほ：ほうば：バトンび：病院ぶ：
ＢＯＯＫ・ＯＦＦべ：別ぼ：ぼくぱ：パラダイスぴ：ピアノぷ：プレゼントぺ：ページぽ：ポイントま：満足み：みたりむ：むーっちゃめ：
迷惑も：戻ってや：やっぱゆ：ゆっくりよ：予約ら：ラブラブり：リュックる：ルートれ：連絡ろ：６月わ：罠を：[をな、] ん：んー

■表 4.3 の評価×
検出した対象語列：に+で+ある
新聞頻度：1 回，ブログ頻度：1 回
品詞：助詞-格助詞-一般+助動詞+助動詞
例文：
って言ってて、そう思ってみると、社会不適格さもエロさもけだるさも、主人公を男にしたら、どこ [にである] 小説になってしまうかな。

図 4.1 評価×の事例

第5章　擬似的に作成したデータを用いた実験

5.1　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 3 章の提案手法で検出できるかを調べる．提案手法の手順は 3 章の通りであるが，4 章とは違い，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 (3 章のとおり検出には文書 A の頻度 0 を利用する).

5.2　新聞に 1,000 文のブログ文書を混ぜた実験

文書 A,B に 4 章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 5.3 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 5.1 のようになる．4 章と同様に χ^2 検定を行うのだが，ただし今回は c と d が「4」以下なので，検定統計量の計算式が 4.1 式と異なる．Yates(イエーツ) の補正式

$$T = \frac{n * (|a * d - b * c| - \frac{n}{2})^2}{e * f * g * h} \tag{5.1}$$

を用いる．この式を用いて検定統計量を求めると，表 5.1 は $T = 27.68 > 3.84$ なので，混ぜた文と，もとの文とは独立でない (関係がある) ことになる．また，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．

表 5.1 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

5.3 ブログに 1,000 文の新聞の文書を混ぜた実験

文書 A,B に 4 章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 5.2 のようになる．同様に χ^2 検定を行うことにより $T = 8.40 > 3.84$ となり，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．

表 5.2 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

5.4 考察

5.2 節と 5.3 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 (5.3 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 5.3 に示す．これは表 5.2 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．

表 5.3 5.3 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

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 5.4 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 5.4 5.3 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

次に 5.2 節と 5.3 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 5.5 に示す．表 5.5 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．また適合率については，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

表 5.5 5.1 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
5.2	0.003	0.273	0.006	0.006	0.018	0.016
5.3	0.002	0.200	0.004	0.007	0.021	0.011

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，5.2 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 5.6 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 5.5 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．

表 5.6 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%)

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．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．

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．

第6章　追加実験

5 章の結果をもとに，実験条件に修正を加え，追加の実験を行う．なお，5 章より，新聞にブログを混ぜた結果の方が，ブログに新聞を混ぜた結果よりも検出数が多く，わかりやすいため，追加実験は新聞にブログを混ぜた場合のみを行うこととする．

修正点は以下の通りである．

- 修正対象の文書を，新聞の経済面のみとする
- 新聞での頻度を,「0 回」から「1 回以下」とする

1 つ目の修正では，新聞記事は紙面によって口語的な表現が使われていることが分かったので，それを考慮する狙いがある．つまり経済面が他の紙面に比べて，堅い表現で構成されているという点を利用する．

2 つ目の修正では，1 つ目の修正を行っても，少なからず口語的文が新聞に混ざっていることを想定した．経営者や閣僚のコメントやインタビューが予想されるため，新聞記事で 1 回は出現を認めることとする．

6.1　実験結果

先述の修正を加えて再実験をした結果を表 6.1 に示す．

ここでベースラインは $fr_b \geqq 0$ であり，つまりブログ頻度を用いない方法である．また提案手法はブログ頻度を用いるものであり，ベースラインの frb $\geqq$ 0 の場合以外である，frb $\geqq$ 1，frb $\geqq$ 2, ..., frb $\geqq$ 10 が相当する．

5 章の結果に比べ，出力数が多くなり，また，誤りを正しく検出できた数も増えていることがわかる．また，表 6.1 ではブログ頻度が増えれば増えるほど適合率が上昇していることがわかる．これは 4 章の人手評価と同じであることから，新聞において，ブログ頻度が増えるほど文体が変わっている (好ましくない) 率が高くなっていることを再度確認できたことになる．

表 6.1 再実験の結果

ブログ頻度 (fr_b)	誤りを正しく検出できた数 (文)	総検出数 (文)	適合率
$fr_b \geqq 0$	381	1275	0.299
$fr_b \geqq 1$	301	875	0.344
$fr_b \geqq 2$	276	733	0.377
$fr_b \geqq 3$	265	652	0.406
$fr_b \geqq 4$	260	616	0.422
:	:	:	:
$fr_b \geqq 10$	227	485	0.468

6.2 検出結果の例

ここで，うまく処理できた例を示す．表 6.2 は混ぜたブログ文を検出できた例である．今回の結果では，語尾に口語的な表現が多かった．また，ブログ記事のくだけた表現の特徴として，以下のような表現が多かった．

- 感動詞の多用，過剰な連続
- 句読点の多用，過剰な連続
- 語尾に「ね」や「よ」等の終助詞を多用

次に，提案手法でうまく新聞記事の表現を検出しなかった例を載せる (表 6.3)．ブログ頻度が低いために提案手法では誤りとせず，正しく処理することができた．提案手法の有効性がうかがえる．ベースラインではブログ頻度を利用しないために新聞の頻度が少ないだけで誤りとしてしまう．ベースラインではこれらの表現は誤りとして検出してしまっていた．提案手法はベースラインより有効であることがこのことからも言える．

表 6.2 ブログを検出できた例

対象語列	fr_n	fr_b	使われている文
なー	0	3	楽しかったなー
じゃ+なかっ+た+。	0	3	「Ｍａｙ’ｎ☆Ｓｐａｃｅ」って、ＣＤで聴いたときは、あまり好きじゃなかった。
う+か+なぁ	0	3	前編＆後編セットの前売り券を買ってしまおうかなぁ☆
ねん+。	0	4	クロー：え～とね、本を読んでしまうとね、夢中になっちゃうねん。
ましょ	0	6	土曜の午後はＦＡＩＲＧＲＯＵＮＤｃａｆｅでハワイアンーな気分になりましょ
っす+。	0	9	あ゛ーーーテストいやぁぁ＞＜;;だるいっす。
なかっ+た+けど+、	0	10	衣装もダンスも全然見えなかったけど、ライブならではの音と、生歌に感動モン。
ませ+ん+よ	0	11	裏じゃありませんよ（笑）
ねぇ	0	17	部長の姿みえねぇ～～～。
あ～	0	20	あ～～ダメっ嬉しすぎて小鼻が膨らんじゃうぞっっ（＊≧ｍ≦＊）
なぁ+。	0	21	素人って独創的なことするなぁ。
まし+た+よ+。	0	25	そんなお疲れモードな相方を迎えに宇治の平等院まで行ってきましたよ。
です+けど+ね+。	0	29	まぁ～所詮たらればですけどね。
いろんな	0	28	いろんなことがありすぎました。
やら	1	33	ダラダラと汗をかき、背中やら喉やらを流れる不快感
だっ+たり	0	31	自衛隊ではね、内務って言って、洗濯だったりアイロンだったり整理整頓だったり要は自分の周り、何から何までピカピカ・きっちり整えないといけないらしい。
まぁ+、	0	44	まぁ、野郎も俺とアームと酒浴びましょうや！
です+よ+ね+。	1	69	みんなタラバガニ科に属する蟹の王様キングクラブでタラバガニの仲間なんですよね。

表 6.3 新聞を検出しなかった例

対象語列	fr_n	fr_b	使われている文
ところが、一方で、	0	0	ところが、一方で、短時間で燃焼させると不完全燃焼が起きやすく、粒子状物質が増える結果を招く
だっ+た+と+の	0	0	米国でも訪日は失敗だったとの評論が多いようだ。
ものの+、+この	0	0	米景気回復の先行きに疑問の余地が残っていることから、高値警戒感も台頭してきてはいるものの、この株高は低金利と大統領選を控えたブッシュ政権の積極的な経済政策に対する期待を反映したもので、しばらくは続くとの観測が強い。
また+、+単なる	0	0	また、単なる職探しなどは認めないとされた労働移動についても、外国企業が建設工事を請け負った場合、その企業との契約で来日する単純労働者の受け入れなどは二国間協議の対象となっている。
た+ばかり+だ+が+、	0	0	一月のブッシュ大統領来日時に日本メーカーは完成車輸入を増やす用意があることを伝えたばかりだが、道のりは困難なことが浮き彫りになっている。
として+、+どの	0	0	コメ関税化の行方も含め、交渉の先行きは米国が早期合意の推進力として、どのような指導力を発揮するかにかかってきた。
から+に+は+、	0	0	米大統領がビッグスリー首脳と来日するからには、何とかしたいと真剣に考えている。
など+も+この	0	0	ドンケル・ガット事務局長は、最終合意案の修正に消極的で、米国や豪州などもこの方針に同調している。
しかも+、+この	0	0	しかも、この長時間労働が日本人の生活から、ゆとりを奪っているだけでなく、海外からは日本企業の競争力を高める原因になっているとの批判も多い。
ばかり+で+なく	0	0	コメばかりでなく菓子でも抵抗しては、日本が新ラウンドつぶしの張本人にされかねず、農水省はまたも譲歩を迫られた。

6.3 追加実験に対する考察

提案手法は，誤りを正しく検出できた数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ため，ベースラインよりも効果的な利用方法が考えられる．例えば，文書B(ここではブログ) 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，誤りを正しく検出できた数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することができる．

第7章　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．

本研究では，修正対象の文書を，新聞の経済面のみとした場合の疑似データに基づく実験も行った．この実験では新聞データにブログの文を混ぜておき，新聞データからブログデータを検出できればブログのようなくだけた文を検出できたという意味で検出に成功したと考える．その実験において，複数の文書の頻度を用いないベースラインの方法 (ブログでの頻度を用いない方法) では，ブログ文の検出は 0.299 の適合率であった．また，複数の文書の頻度を用いる提案手法 (ブログでの頻度を用いる方法) では，ブログでの頻度が 1 以上，2 以上，10 以上とした場合にそれぞれ 0.344，0.377，0.468 という適合率を得た．提案手法はベースラインよりも高い適合率であった．提案手法はブログでの頻度が高い場合ほど高い適合率を得ることを確認した．

ブログでの頻度が高い場合ほど高い適合率を得ることができるので，提案手法は効果的な利用方法が考えられる．例えば，ブログでの頻度が高い表現から修正候補として提示することで，誤りの可能性が高い表現から人手でチェックしていくことが可能である．

本研究では，新聞とブログのデータしか用いなかった．文書としては，論文やウィキペディアなどを利用することも考えられる．これらの文書に対して提案手法を利用することは今後の課題とする．

謝辞

本研究を進めるに当たり，終始に渡り研究の進め方や本論文の書き方など，細部にわたる御指導を頂きました，鳥取大学工学部知能情報工学科計算機工学講座 C の村田真樹教授に心から御礼申し上げます．また，本研究を進めるにあたり，御指導，御助言を頂きました，村上仁一准教授，徳久雅人講師に心から御礼申し上げます．その他様々な場面で御助言を頂いた計算機工学講座 C 研究室の皆様に感謝の意を表します．

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/