Wikipediaの編集履歴を用いた書き換えパターンの抽出

金山 博　　荻野 紫穂
日本アイ・ビー・エム株式会社　東京基礎研究所
{hkana, shiho}@.jp.ibm.com

1　はじめに

ビジネス文書や仕様書など実社会で使用される文書に関して,「校正支援システムなどを用いて文書の質を上げたい」という要望は大きい．しかし，その概念的な要望が同じであっても，期待される具体的な修正項目は，その文書の分野や利用目的等によって異なる．また，一つの文書の中にも,「文法のエラーを直したい」「プロジェクト固有の用語やフォーマットに統一したい」「決定事項の内容変更に伴う修正を行いたい」といったように，異なる目的に由来する修正箇所が混在する．

このため，人手で規則を書く場合でも，修正前の文章と修正後の文章とを比較して自動または半自動で規則を獲得する場合でも，ある分野で獲得・使用した修正規則や修正例を別分野に適用するには，個々の修正規則や修正例について，規則を適用する分野の目的に沿っているものを分類して抽出する必要がある．

我々の研究は，文書の修正前後の差分から，文書の品質向上のための書き換え規則を自動抽出すると共に，その規則が他の分野にも適用できるかどうかなどの属性を含む，書き換え規則のコンテクストを付与することを目標とする．

本稿では，その一環として，Wikipedia[1]の編集履歴をデータとして用いる．Wikipediaの更新には，情報の詳細化や正確化など内容に関わるものや，書き誤りの修正や用語の統一などの表記に関するものなど，背後には様々な意図がある．集合知の特徴として，Wikipediaの文書の更新はほとんどが文書を何らかの意味で改善しているものという仮定のもと，本稿では，文法・語句の修正および表記統一のための，概ね内容を変更しないような修正に着目し，それらを同定する．さらに，Wikipediaを含めた一般の文書の修正に適用できるように，典型的な修正を文字列変換の規則の形式で抽出する．

それらの有効性を示すために，内容の変更と表記の修正の機械学習を用いて分類する実験と，Wikipediaの最新版の中における文字列の頻度を用いて書き換えを推奨すべき文字列を判定する実験を行った．それらの規則を文書に適用することによって，一般的な文書校正作業において有用であるかという観点で評価する．

[1] http://ja.wikipedia.org/

2　関連研究

分野や目的ごとの基準に従って文書の校正支援規則を定義する手法は，非母国語話者の文章校正の研究や要求仕様書の処理など，様々な分野で行われている．例えば，[4]には，IEEE Std. 830中の要求仕様書が守るべき項目に沿って校正支援の警告規則を定義した例が挙げられている．これらの規則の中には，例えば「適当な」という形容動詞に対する警告のように，要求仕様書という分野においては重要な指摘だが，別の分野では必要でないものも含まれる．こうした例から，修正規則には他分野に適用できるものとできないものとがあり，ある分野で使用された修正規則や修正例を他分野で活用するには，新たな分野に合った修正規則を選択する必要があることがわかる．

非母語話者の文書校正のために，松崎らは，校正の前後の英文を比較して規則を抽出している[3]．Wikipediaの編集履歴を扱うとなると，修正の前後が文法的な修正の結果の変化とは限らないことと，修正後の状態が文法的に正しいという仮定が成り立たないことから，本稿では別の工夫が必要となる．その他，荒牧らは，タイピングの特性に踏み込んで典型的な誤りについて考察している[2]．

Wikipediaの修正履歴のデータは，記事の信憑性の判定に有用であることが知られている．Adlerらは，ある書き手が加えた変更がどれだけ長く残存するかを見ることにより，書き手の信頼性を算出する手法を提案した[1]．さらに，鈴木らは日本語のデータにおいても実験を行い，不適切な編集を行う著者の傾向を捉えられたと報告している[5]．

3　Wikipediaの変更履歴

3.1　差分の抽出

今回は，Wikipediaの2010年1月版の変更履歴のデータを用いた．<page>タグで囲まれた各記事の中に，<revision>タグで囲まれた各時点での編集結果(「版」と呼ばれる）がMediaWiki形式で保存されている．対象のデータの中には，特殊ページを含めて約168万記事，約2849万の版がある．同一記事の隣接する版 (第 $n-1$ 版と第 n 版) が同数の文からなる場合に，それぞれの文を比較して，相異点があるものの

表 1: Wikipedia の修正の例
上段，下段がそれぞれ修正前後の文．

a	所在地 = [[東京都]] [[港区]] 所在地 = [[東京都]][[港区（東京都）|港区]]
b	{{Category:三国志の登場人物|ぼうちゅう }} {{Category:三国志の登場人物|ほうちゆう }}
c	=== 音楽 CD === === 音楽 CD・ドラマ CD ===
d	建設中。2010 年完成予定 建設凍結中
e	県中央部に位置する湯布院町の由布院温泉、… 県中央部に位置する由布市の由布院温泉、…
f	[[コンピューター将棋]] [[コンピュータ将棋]]
g	癌にかかっていることが、記者会見で明らかにした。 癌にかかっていることを、記者会見で明らかにした。

うち，文長の差が 20 文字以内の約 1765 万ペアを抽出した．なお，2 版以内に逆方向の修正がされているペアは除外した．これは，明らかに悪戯と思われる追記などによるノイズを減らすためである．

表 1 は変更点の例である．それぞれのデータには，表に示した修正前後の文のほか，編集者，タイムスタンプ，編集時のコメントが付与されている．

3.2 修正の意図

表 1 にあるような例を見ると，Wikipedia の修正点は，主に以下の 3 つに分類できる．

形式修正 MediaWiki の形式に特有のもの

表記修正 文法や語句の誤りの修正・表記の統一など，内容を変更しないもの

内容修正 上記に当てはまらない，主に内容自体の変更

“形式修正” は，表 1 の a, b のような，リンクの追加や特殊な読みの表記の修正などを指す．これらは，自然言語よりは文書形式に依存するものなので，本稿での議論の対象外とする．残りの修正のうち，“表記修正” が，本研究で主に扱うものである．記事自体の内容を変えないことから，別の文書であっても，修正前の表記と同じものが見つかったら，それを修正後の表記に変えることを提示できる性質を持つ．主なものは表 1 の f のようなカタカナ語の表記統一，g のような文法誤りの修正である．“内容修正” は，それに当てはまらない，記事そのものの書き換えであり，編集履歴の修正点の大半を占める．

一方で，表記修正と内容修正の境界は曖昧である．表 1 の e にあるような自治体名の公的な変更に伴う修正の場合，文の内容を変更しているが，同様の表記をがあった時に書き改めるよう指摘することは，一般的な校正規則として有用な場面が多い．本稿で対象外にしている他の分類についても，多くの修正がなされている現象とその文脈を同定できれば，再利用可能な知識となり得るものが多い．

表 2: 編集前後の差分として頻出する文字列の例
ϕ は空文字列を指す．

修正前	修正後	頻度
ϕ	、	7,751
ϕ	。	7,228
、	ϕ	6,223
。	$\phi\phi$	5,490
の	$\phi\phi$	4,502
ϕ	の	3,864
予定	ϕ	3,699
・	$\phi\phi$	3,644
ー	$\phi\phi$	3,390
ϕ	に	2,800
る	た	2,735
ϕ	を	2,288
ϕ	と	2,184
ϕ	い	1,984

修正前	修正後	頻度
ϕ	日本の	1,749
⋮	⋮	⋮
の	に	1,189
る	った	1,187
⋮	⋮	⋮
トッ	ド	1,118
ー	ϕ イ	1,113
に	の	1,029
⋮	⋮	⋮
初	始	694
ϕ	等学	656

3.3 文字列の変更部分

次に，頻繁に書き換えが行われる現象を見るために，２文の間で変更されている部分の文字列を抽出し，その頻度を観察した．なお，ここでは，同一のユーザーが大量に同じ書き換えをしていることによる頻度の増加の影響を軽減するため，ユーザーの異なり数をもって頻度とする．高頻度の書き換えのうち主なものを表 2 に示す．

この結果から，頻繁に修正されている文字列の中に，格助詞を追加・削除・変更するものや，カタカナ語の表記を変更するもの,「る→た」「予定→ ϕ」のように時間の変化に対応させたもの，など特徴のある修正を見いだすことができる．また，書き換えが双方向に行われていて，一方が望ましいとはいえないものもある．どの現象をとっても，変更された文字列の部分だけで 3.2 節のような意図の分類をしたり，自動的に書き換えを行うことは困難である．次節にて，これらのデータを再利用可能な知識に変換する方法について述べる．

4 意図の分類と規則の抽出

再利用可能な規則を抽出するためには，一般性の高い修正を同定することと，その書き換えが適用できる最適な条件を求める必要がある．本節では，この 2 点の手法について述べる．

まず，3.2 節に挙げた 3 つの分類のうち，形式修正については，予め以下の規則を用いて除外しておく．

- リンクの追加・削除に関するもの（[[ ]] や | などが修正点となっている場合 – 例：表 1 の a）
- Category リンクの中の特殊な読み仮名表記（例：表 1 の b）
- タグの追加や削除（{{Food-stub}} のような Wikipedia で定義されたタグ）
- 半角・全角文字の変換や，機種依存文字の除去

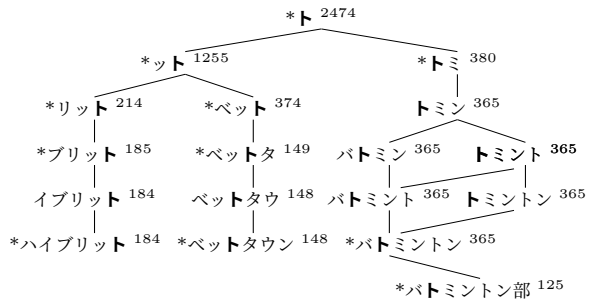

図 1: 「ト→ド」の修正が行われる文脈の半順序関係

表記修正と内容修正を区別するために，表記修正に相当する事象を予め定めたのでは，校正処理等に使える新たな事象を獲得することができない．そこで，変更された部分の文字種や文法的属性を素性として機械学習でこれらの二値分類を行う．さらに，ユーザーの情報や，更新の間隔，コメントなど編集履歴に付随する情報もそれが内容の変更であるか否かの判定に寄与すると考えられるため，これらを素性に加える．5.1 節にて，具体的な素性と実験結果を示す．

次に，高頻度で書き換えが行われている事象の中から，自動的な書き換えの推奨が可能になるような規則を生成する．3.3 節で見たような，変更された文字列だけでは，実際に書き換えを行うべき状況を同定することは困難である．特に，書き換えの左辺が ϕ の場合は，当然ながら何らかの文脈が必要となる．

難しいところは，Wikipedia の編集履歴は，いわば修正の「正例」である．新しい版を作る際に編集をしなかった事象を「負例」と捉えることもできるが，修正すべき点であっても見過ごされている場合がほとんどであることから，正例が極端に疎になる．また，特にカタカナ語の書き誤りなどの場合，書かれている語の単位が予め辞書に登録されている形態素の単位と一致しないため，形態素解析をすれば文脈が求まるというわけでもない．

そこで，修正が行われた文脈の頻度を集計する．図 4 は,「ト」を「ド」に書き換えられた場合に左右の n 文字を文脈とした場合の，文脈の詳細さの半順序関係を図示したものである．数字は修正の頻度を表し，隣り合うノードの間で頻度が等しいもののうち最大の文脈となるもの（図 4 で '*' が付いたもの）を “極大文脈” とし，それらのうち頻度 f_c が s 以上のものを，条件の候補として考える．さらに，Wikipedia の最新の版における，文脈の文字列（例：「ハイブリット」）の頻度 f_b とそれを修正した文字列（例：「ハイブリッド」）の頻度 f_a を用いて，修正がどれだけ必須かを示す指標として

$$score(c) = \frac{f_c \log f_a}{f_b}$$

を導入する．これは，同様の修正が頻繁であり，かつ最新の記事で多く使われるものに高い値を与えるもので，この score が高いものから適切な文脈を抜き出す．その結果は 5.2 節に示す．

表 3: 分類の際に用いる素性集合

素性集合	素性
S	変更部分（修正前）の長さ 変更部分（修正後）の長さ 変更部分（修正前）の文字種 変更部分（修正後）の文字種 変更部分（修正前）の文字列 変更部分（修正後）の文字列 変更部分（修正前）の最初の文字 変更部分（修正後）の最初の文字 変更部分の長さの変化の有無 変更部分の直前の文字の文字種 変更部分の直後の文字の文字種 修正前の文の長さ 修正後の文の長さ 変更部分（修正前）に拘わる形態素の数 変更部分（修正後）に拘わる形態素の数 変更部分に拘わる形態素の数の変化の有無
W	前の版とユーザー名が一致するか 前の版と同日に書かれているか 前の版から 1 時間以内に書かれているか コメントに「typo」の文字が含まれているか ユーザー名の文字列

表 4: SVM を用いて表記修正を同定する性能

素性集合	Precision	Recall	F1
S	81.2%	83.4%	82.3%
S+W	81.4%	82.5%	81.9%

5 実験

本節では，3.1 節に示した Wikipedia の編集履歴のデータを用いて行った実験の結果を示す．

5.1 表記修正と内容修正の自動分類

修正点を表記修正と内容修正に分類するために，修正前後のペアから形式修正と判定されたものを除いたものから無作為に抽出した 1,479 件に，いずれのクラスに属するかのラベルを人手で付与したところ，表記修正，内容修正はそれぞれ 145 件，1,334 件と，表記修正と認められるものは 1 割未満だった．

これをサポートベクタマシンを用いて 2 値分類する．素性集合として，表 3 にあるように，文字列に関するもの (S)，Wikipedia 特有の情報 (W) を導入した．

10-fold cross validation の平均値を表 4 に示す．結果として，W に属する各素性は一定の効果が認められたものの，全体の性能は W を導入しない時の方が高かった．但し，両者の間に有意な差は認められていない．このことから，内容の修正をしているかどうかの判定にとって，Wikipedia 特有のメタデータは必須ではないといえる．

recall 低下の原因，すなわち表記修正といえるものが内容修正と判断されてしまう場合には，主に同様の例が学習データに現れない漢字表記の修正（「取り」→「トリ」など）が目立った．漢字の読みのデータを

表 5: 修正の文脈の頻度と score

「ト→ド」の修正の文脈

文脈	f_c	f_b	f_a	score
バ**ト**ミントン	365	5	2,343	250.0
ベッ**ト**タウン	148	7	893	62.38
ダブルベッ**ト**	90	4	46	37.41
ベッ**ト**タ	149	15	964	29.64
ハイブリッ**ト**	184	60	3,103	10.71
ツ＝ハラル**ト**・フレンツ	60	11	71	10.10
ベッ**ト**ルーム	87	21	89	8.076
のベッ**ト**	88	59	649	4.194
レオパル**ト** 1	19	9	53	3.640
ビッ**ト**ソン	33	25	442	3.492
⋮	⋮	⋮	⋮	⋮
ブリッ**ト**	185	659	3,197	.9839
⋮	⋮	⋮	⋮	⋮
ヘッ**ト**	11	168	15,356	.2705

「の→に」の修正の文脈

文脈	f_c	f_b	f_a	score
のよって	50	46	195,025	5.750
のおいて	23	38	153,664	3.139
のなっ	13	41	266,611	1.720
の属し	17	43	6,750	1.514
のよる	46	232	242,441	1.068
⋮	⋮	⋮	⋮	⋮
の登場する	14	468	24,976	.1315
⋮	⋮	⋮	⋮	⋮
的**の**	16	4,878	248,954	.0177

用いれば改善できるであろうが，未知の書き誤りの例を探るためには，そういったリソースを用いないほうが望ましいともいえる．

逆に，precision 低下の原因は,「直流電化」→「交流電化」のような，同種の一文字の置換によって内容の修正をしているものが，表記修正と判定される場合が多い．いずれにせよ，大幅な改善の余地がある．

5.2 規則の抽出

次に，高頻度で認められる修正履歴から，一般的に適用できるような書き換え規則を抽出する．すなわち，表 2 にあるような典型的な書き換えが成り立つための条件を，周囲の文脈を用いて求める．ここでは,「ト→ド」と「の→に」という書き換えが適用できる文脈について実験を行う．

まず，実験 1 の分類器を適用し，表記修正と推定されるものに絞ってから，$n = 5$ 文字の前後の文脈を見て，4 節で導入した極大文脈の修正のうち，$s = 10$ 以上の頻度を持つものに対して，同じく 4 節の score を計算し，その閾値を調整しながら規則の抽出を行った．

表 5 が獲得された規則の例である．これらの閾値を 5.0〜0.1 まで動かした時に Wikipedia の最新版の文にマッチした数と，そこから 100 件ずつを任意抽出した時にそれらが修正すべきであるかどうかを人手で評価した正解率，および正しい指摘の推定数を表 6 に示す．これらより，導入したスコアが異なるタイプの誤りに使える指標であること，高い閾値を設定すれば正確な修正点の検出ができること，人手での確認に使える時間に応じて検出できる現象が増やせることがわかった．また，Wikipedia では常に修正が行われているとはいえ，最新の版にも書き誤りが残っていることもわかる．

表 6: 閾値別の，発見される修正点の数と正解率

閾値	指摘数	正解率	推定正解数
5.0	184	96%	177
3.0	324	88%	285
1.0	1,365	73%	996
0.5	4,942	32%	1,581
0.1	23,451	10%	2,345

6 おわりに

本研究では，Wikipedia の編集の履歴から一般的な校正処理に使う規則を抽出し，Wikipedia の修正ができるという，文書の「自浄」ができることを確認した．文脈の捉え方は前後の文字列を見るだけだが，これに係り受けや他の条件などを加えて，より適切な条件付けをして，多様な校正を可能としたり，文書を扱う場面に応じて，構文的な修正に留まらず，意味的な修正を含めて文書そのものだけでなく，文書の変更というメタな知識を共有することによって文書作成や知識共有の効率化に繋げていくことが期待できる．

参考文献

[1] B. Thomas Adler and Luca de Alfaro. A content-driven reputation system for the wikipedia. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pp. 261–270, 2007.

[2] 荒牧英治, 宇野良子, 岡瑞起. Typo writer: ヒトはどのように打ち間違えるのか？言語処理学会第 16 回年次大会論文集, 2010.

[3] 松崎幸太郎, 田中久美子. 英文とその校正からの書き換えルール自動抽出. 言語処理学会年次大会論文集, 2006.

[4] 竹内広宜, 荻野紫穂, 中田武男, 坂本佳史, 福岡直明. テキスト分析技術を用いた開発関連文書の品質分析. 組み込みシステムシンポジウム, 2009.

[5] 鈴木優, 金本径卓, 川越恭二. Wikipedia の編集履歴を用いた記事の信頼性導出. 人工知能学会第 20 回セマンティックウェブとオントロジー研究会, 2009.