7Q-02

ソーシャルメディアに現れる くだけた表現を含む口語的表現の自動修正方式

星野 恵以子† 寺田 篤史† 村上 久† 秋吉 政徳†
神奈川大学†

1. はじめに

近年、有益な情報が発信されているソーシャルメディアを対象としたデータマイニングが注目されている。この際に、「ぉはよー」や「わヵらない」などのくだけた表現が多数含まれ、一般の形態素解析器では未知語として検出されるため十分な解析結果を得られない。こうした問題に対し、従来では、くだけた表現を人手により辞書登録することで解決しようとされているが、その全てを登録することは困難かつ人的コストが大きいことが課題に挙げられる。

池田ら [1] は、自動的にくだけた表現を一般的な表現に修正することで高精度な形態素解析を実現する手法を提案している。この際にくだけた表現とは、表 1 に示すように一般的な表現から予め定めた派生表現をもとにくだけた表現を定義している。しかし、くだけた表現は絶えず増え続けているため、修正されることのない表現も存在する。さらに、表 2 に示すような形態素解析によって未知語が判断されるものには一般的な表現も含まれるが、その判断をしておらず、修正対象となってしまうことがある。

本稿では、形態素解析によって未知語と判断された語が、(i) 一般的な語であるが未だ辞書に登録されていない語であるのか、(ii) くだけた表現であるのか、を自動的にまず区別する。これによって既存研究で定義されたくだけた表現に加え、新たに使われ始めているくだけた表現を修正することができる。さらに本来修正されるべきではない箇所の誤修正を防ぐための手法を提案する。

表 1: 既存研究で定義されるくだけた表現

分類	派生表現	一般的な表現
形状が似ている文字の代替	ぉはよぅ わた U	おはよう わたし
発音記号の変化傾向に合わせた表記	すっごーい どしたの？	すごい どうしたの？
カタカナの代用	カッコイイ	かっこいい

表 2:未知語として検出された一般的な単語

分類	例
製品名、商品名	Mac, Windows, プレイステーション
新語、造語	IoT, スマホ、ポケモン

Automatic correction method for colloquial expression including inappropriate expressions in social media
† Eiko Hoshino, Atsushi Terada, Hisashi Murakami, Masanori Akiyoshi, Kanagawa University

2. くだけた表現の定義

本研究におけるくだけた表現は、表 3 に示すように本来の表記では形態素解析辞書に登録されているが、表記上の揺らぎによって未知語として検出されるものと定義する。くだけた表現は、移り変わりが激しく絶えず増え続けるため、表 3 に挙げた例以外の表記にも対応させる必要がある。

表 3：増え続けるくだけた表現

分類	表記例	本来の表記
英語を日本語で表記	おっけー	OK
略した言い回し	りょ	了解
余計な言葉の付け足し	つらたん	つらい
. .	. .	. .

3. 提案手法

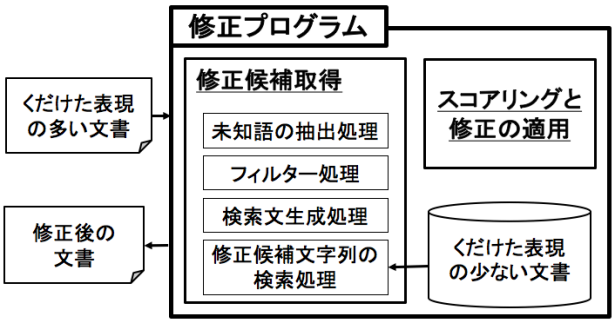

図 1: 自動修正方式の構成

提案手法の全体像を図 1に示す。本提案手法では、くだけた表現の多い文書として、Twitter に投稿されたツイートを修正対象とする。処理の流れは大きく分けて、未知語検出箇所の “修正候補取得” とその修正候補の “スコアリングと修正の適用” がある。

3.1 修正候補取得

3.1.1 未知語の抽出処理

修正候補の取得は、まず、ツイートから未知語の抽出を行う。くだけた表現の多くは、形態素解析辞書に登録されていないため、未知語として検出される。抽出処理だけでは、一般的な言葉が未知語として検出され、修正されることがある。例えば、「この ポケモン強い」という文章で、「ポケモン」は一般的な言葉にも関わらず、

形態素解析辞書に登録されて いないため、未知語と検出されてしまう。この場合、「強い」という言葉に影響され、「この力強い」とい う文章に修正されてしまう。そのため、検出された 未知語に対してフィルター処理を行う。

3.1.2 フィルター処理

フィルター処理では、検出された未知語がくだけ た表現なのか、あるいは一般的な言葉であるかを区 別する。未知語に対して、TF-IDF(Term Frequency- Inverse Document Frequency) 処理を行い、TF-IDF 値が設けた閾値よりも高い数値となった場合はくだ けた表現、低い数値となった場合は一般的な言葉と 判断する。表 4 に例を示す。表 4 中の“ぉはよーご ざいます” というツイートの未知語“ぉはよー” はく だけた表現、“リツイートしてくださった皆様ありが とうございます” というツイートの未知語“リツイー ト” は一般的な言葉として判断される。

表 4：TF-IDF 処理例

ツイート	未知語	TF-IDF 値
ぉはよーございます	ぉはよー	1.82
リツイートしてくださった皆様ありがとうございます	リツイート	0.53

TF-IDF 処理の元となるデータとして Twitter 上 に実際に投稿されたツイート群を用いることで、投 稿時期に主流なくだけた表現に対して特徴量を算出 することができるため、くだけた表現の移り変わり に対応することができると考えられる。

3.1.3 修正候補文字列の検索処理

フィルター処理を行った結果、くだけた表現であ ると判断された未知語に対して、出版書などの校正 がなされているくだけた表現の少ない文書に検索を かけ、修正候補文字列を取得する。本研究ではくだ けた表現の少ない文書には、国立国語研究所が提供 する日本語話し言葉コーパス [2] と文庫コーパス [3] を用いている。

3.2 スコアリングと修正の適用

スコアリングには、(1) 修正候補文字列の出現頻度、(2) 修正文字列間の編集距離を用い、これらをもとに 総合スコアを算出する。その後、修正候補文字列の総 合スコアが最も高いものを修正文字列とし、ツイー トの修正を行う。

4. 実験

4.1 予備実験

提案手法のフィルター処理で設定した閾値による 区別の精度を検証する。処理を行うための元データ として 12000 文のツイートを用いる。まず始めに仮 の閾値を設けて未知語の区別を行い、正しい区別が 行われているかの確認を人手で行う。その後、始めの閾値とは異なる閾値を設けて区別を行う。

表5：設定した閾値による区別精度

	総数	正	負	適合率	再現率	F値
くだけた表現	1515	1207	308	0.796	0.658	0.696
一般的な言葉	4674	4051	625	0.866	0.929	0.896

この処理を繰り返し行い、最も区別の精度が高かった閾値 をフィルター処理に用いる閾値とする。評価実験の結果、閾値を 0.75 に設定することとした。表 5 に算 出した閾値による区別の精度を示す。

4.2 評価実験

Twitter 上に実際に投稿されたツイート 500 文に対して、既存手法と提案手法によるツイートの修正 を行い、表 6 に示す評価指標にて人手で正しい修正を行うことができるかの評価実験を行う。

表6：修正後の文章に対する評価指数

判定	条件
○	文章の構造及び意味的ともに正しい修正がなされている
△	文章の構造あるいは意味的において、どちらかは正しい修正がなされている
×	文章の構造及び意味的ともに正しい修正がなされていない

4.3 結果

表 7 は、表 6 の評価指標に基づき、既存手法と提案手法においてツイートの修正を行った総評価結果を示している。

表7：各方式における評価結果

方式	判定			合計
	○or△	○	×	
		△		
既存手法	426	333	74	500
		93		
提案手法	459	357	41	500
		102		

結果として、既存手法では「◯or△」と評価したツ イートは 500 文中の 85.2%、提案手法では 91.8%と なり、提案手法の方が既存手法に比べ、正しい修正 がなされるツイート数が上昇していた。

5. おわりに

本稿では、実際に投稿されたツイートを対象に、ツイート中に含まれるくだけた表現を自動修正し、本 来修正されるべきではない箇所の誤修正を防ぐ手法 の提案を行った。

参考文献

[1] 池田和文,柳原正,松本一則,滝嶋康弘,“くだ けた表現を解析するための正規化ルール自動生成 手法”, 情報処理学会論文誌, pp.68–77 (2010)
[2] 国立日本国語研究所,“日本語話し言葉コーパス”
[3] 青空文庫,“『青空文庫』パッケージ”