2017 年度卒業研究概要

自由記述型アンケートの分析効率化を図る校正支援システム

大谷研究室
1472017 大沼 龍斗

1. 研究の背景・目的

アンケートデータの自由入力欄には表記揺れ，誤字脱字が含まれていることがある．表記揺れとは，表記間違いや異なる表記方法のことを指す．表記間違いには，「顕微鏡」という単語の「微」という文字が「徴」になり，「顕徴鏡」と間違って表記される例が挙げられる．また，異なる表記方法には，英単語である「interface」をカタカナで表記する場合に「インターフェース」「インタフェース」「インタフェイス」などの複数の候補が存在する例が挙げられる．アンケートデータに含まれる単語の頻度分布をコンピュータで分析する際には，表記の異なる同じ単語を正しく集計するために，表記揺れや誤字脱字を修正する必要がある．

Microsoft Office Word(以下 word)には文法が間違っている可能性のある言葉を検出して自動的に青色の波線を引く機能や，表記揺れの対象となる言葉を一括で統一する機能などがあるが，表記揺れ・誤字脱字を修正するのはあくまでユーザであり，システムが自動的に修正するわけではない．したがって，word を代表とする表記揺れ・誤字脱字が修正可能なソフトで修正する場合には，大量のデータを修正することは手間であり時間がかかる．

本研究では，アンケートデータの自由入力欄に含まれている単語の頻度分析を効率的に行えるようにすることを目的として，自由記述型アンケートの分析効率化を図る校正支援システムを作成する．

2. システムの概要

本システムは，アンケートデータが記述されたテキストファイルが入力されると，表記揺れ・誤字脱字をそれぞれ修正し，修正されたデータをテキストファイルに記述し新規保存する．本研究では，「スマホ，スマートフォン」のような略語・略称，「飴，あめ，アメ」のような文字の漢字・平仮名・カタカナによる表記違い，「引っ越し，引越し」のような送り仮名の表記違い，「斎藤，斉藤」のような異字体，英単語の「フェイス，フェース」のような表記違い，「ヴィトン，Louis Vuitton」のような表記違い，「ティッシュ，ちり紙」のような呼び名の違い，の 7 つを表記揺れと考える．

表記揺れ・誤字脱字の修正では，まず，オープンソースの形態素解析ツール MeCab を用いて，アンケートに記述された文章を単語に分け，各単語の品詞や活用形などを判別する．次に，解析した文章から単語の品詞や活用形の違いごとに単語を抜き出す．抜き出す対象とする単語は，名詞・固有名詞・動詞・形容詞である．その後，誤字脱字，表記揺れの順に修正する．誤字脱字は，文字列の一致率が最大となる別単語に置き換えることで修正する．文字列の一致率は，比較する 2 単語の文字数のうち大きい方の数値を分母，2 単語で一致する文字数を分子として求め，文字列の一致率が 0.7 以上のときに同一単語と判断する．表記揺れの修正では，始めに，すべての対象単語の同義語を web 上の同義語辞書を用いて検索し，対象単語を含めてグループ化する．次に，同じ単語を含むグループ同士を同一グループと

してまとめる．最後に，グループ内での各対象単語の出現数を比較し，最も出現数の多い対象単語に表記を統一することで表記揺れを修正する．最大出現数の対象単語が複数ある場合はランダムに単語を 1 つ選び，選んだ対象単語に統一する．

3. 評価実験

AI が作曲した楽曲についてのアンケートデータ 63 ファイルを用いて実験した．まず，本研究で作成したシステムによりアンケートデータの表記揺れ・誤字脱字を修正する．その後，修正前と修正後の文章それぞれについて，word の校正機能が表記揺れ・誤字脱字をチェックした箇所と word がチェックしなかった箇所をカウントし，表記揺れ・誤字脱字と考えられる言葉の個数の変化によりシステムを評価した．評価実験結果のうち表記揺れ修正結果を表 1，表記揺れ・誤字脱字の修正前データを図 1，修正後データを図 2 に示す．誤字脱字は「わかりずらかった」「良った」といった単語の修正ができず，すべての箇所で単語が修正されなかった．また，word で検出された表記揺れのうち，「わからず」「分からない」の「わか」が統一されなかったが，「メロディ」「メロディー」，「覚え」「おぼえ」など別の箇所は修正できた．word で非検出だった表記揺れは「AI」「人工知能」，「コラボ」「コラボレーション」だったがいずれも修正できた．さらに，「そう」「みたい」が同義語として判定され，表記揺れ修正前では正しい表記だった箇所が修正後に誤字になった箇所があった．

表 1：表記揺れの修正実験結果

	word で検出	word で非検出
提案手法で修正	8 単語・40 箇所	2 単語・16 箇所
提案手法で非修正	1 単語・4 箇所	0 単語・0 箇所

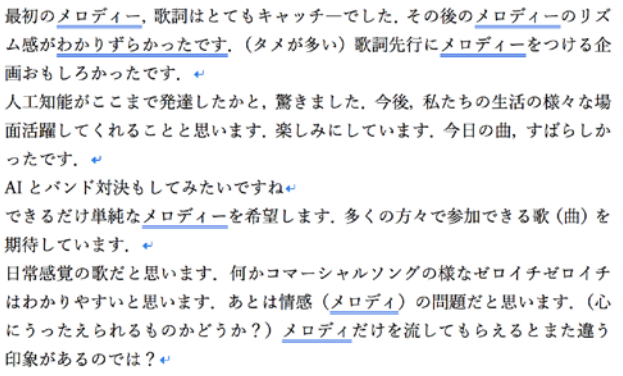
最初のメロディー，歌詞はとてもキャッチーでした．その後のメロディーのリズム感がわかりずらかったです．（タメが多い）歌詞先行にメロディーをつける企画おもしろかったです．
人工知能がここまで発達したかと，驚きました．今後，私たちの生活の様々な場面活躍してくれることと思います．楽しみにしています．今日の曲，すばらしかったです．
AI とバンド対決もしてみたいですね
できるだけ単純なメロディーを希望します．多くの方々で参加できる歌（曲）を期待しています．
日常感覚の歌だと思います．何かコマーシャルソングの様なゼロイチゼロイチはわかりやすいと思います．あとは情感（メロディ）の問題だと思います．（心にうったえられるものかどうか？）メロディだけを流してもらえるとまた違う印象があるのでは？

図 1：校正前のアンケートデータ

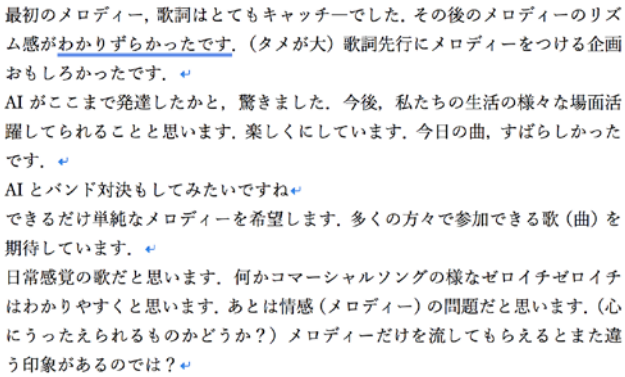
最初のメロディー，歌詞はとてもキャッチーでした．その後のメロディーのリズム感がわかりずらかったです．（タメが大）歌詞先行にメロディーをつける企画おもしろかったです．
AI がここまで発達したかと，驚きました．今後，私たちの生活の様々な場面活躍してられることと思います．楽しくにしています．今日の曲，すばらしかったです．
AI とバンド対決もしてみたいですね
できるだけ単純なメロディーを希望します．多くの方々で参加できる歌（曲）を期待しています．
日常感覚の歌だと思います．何かコマーシャルソングの様なゼロイチゼロイチはわかりやすくと思います．あとは情感（メロディー）の問題だと思います．（心にうったえられるものかどうか？）メロディーだけを流してもらえるとまた違う印象があるのでは？

図 2：校正後のアンケートデータ

4. 考察

評価実験の結果，word が表記揺れと判断しない単語の修正が可能である点や表記揺れの修正精度が高い点などから，表記揺れの修正において本システムの有用性は高いことが示された．しかし，誤字脱字の修正について有用性は示せなかった．誤字脱字が修正されなかった要因として，提案手法では，文字列の一致率が高い単語が存在しない場合単語を修正できないという点が挙げられる．また，表記揺れが修正できなかった単語が存在した要因として，動詞の活用形が異なる場合の同義語検索処理について考慮していないことが挙げられる．本システム全体の課題点は，単語の修正精度が同義語辞書と MeCab の性能に依存することであり，アンケートデータの内容により本システムと異なる辞書を使用する必要がある．現状では複数の問題点が挙げられるが，主に誤字脱字修正における問題点を解決することで，有用性の高いシステムを構築できる可能性がある．