表記・送り仮名による短い文章にも対応できる著者判別

早倉　舞

本研究の目的は，表記と送り仮名の使い方に基づく著者判別手法の有用性の検討である．著者判別は，文章の真贋判定や著者推定を目的として研究されてきた．これまでに様々な手法により著者判別が試みられてきたが，従来の著者判別の手法には対象となる文章が短いほど判別精度が下がる欠点がある．本研究で提案する表記・送り仮名は著者の好みやパソコン等の予測変換により著者の癖が表れやすく，対象の文章の長さに判別精度が左右されにくい．分析により，出版社の影響を抑えた条件下で，従来手法の欠点である判別対象の文章が短いほど精度が低い問題を補えることが分かった．

分析の手順は以下の通りである．まず，著者不明とする文章（以下，著者不明文章）と，比較対象とする文章（以下，比較対象文章）として，著者不明文章を書いた可能性がある著者複数人の文章を同数ずつ用意する．著者不明文章とは，著者が不明なため，著者判別手法によって著者を推定する，と設定する文章である．次に，比較対象文章の長さを固定したまま，著者不明文章を 300 文字から 1 万文字まで変化させ，判別精度，エントロピー，純度の算出を行う．エントロピーと純度はデータを分類できる精度を表し，エントロピーが低いほど，また純度が高いほど，クラスタリングの精度が良いことを表わす．クラスタリングはウォード法で行う．また，判別精度は，クラスタリングの結果著者不明文章に似ている文章上位 5 位以内に著者不明文章と同じ著者の文章が含まれている割合とする．比較対象には，著者判別手法として精度の高い(1)品詞 3-gram による著者判別と，(2)読点の使い方による著者判別を行う．青空文庫の文章のうち新字新仮名のものを分析する．

旧字体の影響を抑えるために，旧字体が新字体に完全に切り替わったとされる 1950 年以降に書かれた 3 著者それぞれ 3 文章，合計 9 文章を分析した．その結果，表記・送り仮名による著者判別はエントロピー0.33，純度 0.78，品詞 3-gram による著者判別はエントロピー0.58，純度 0.67，読点の使い方による著者判別はエントロピー0.42，純度 0.67 となり，表記・送り仮名による著者判別のクラスタリングの精度が品詞 3-gram，読点の使い方よりも良かった．

次に，表記・送り仮名による著者判別の特徴をより詳細に分析するために，年代を問わず集めた 12 著者 63 文章を分析した．その結果，表記・送り仮名による著者判別は，判別精度は著者不明文章の長さが 300～5,000 文字という条件下で 1 割を下回ったが，クラスタリングの精度は品詞 3-gram と同程度である．つまり，表記・送り仮名による著者判別は，判別精度が非常に低いがクラスタリングの精度は品詞 3-gram に並ぶほどに良い．これは，一部の文章は著者の癖とは異なる表記・送り仮名を用いているためだと考えられる．このような例外が生じる原因として，一部の出版社において旧字体や旧仮名遣いの文章を新字新仮名に直す際に表記や送り仮名を変更することが挙げられる．つまり，出版社の編集により著者の表記・送り仮名の癖が充分に反映されていない文章が生じたため，表記・送り仮名による著者判別の精度が低くなったと考えられる．

（指導教員　真栄城　哲也）