CBoW から獲得できる単語ベクトルを用いた小説作品の特徴量定義と著者の推定

竹中 志織（福井大学大学院工学研究科）
黒岩 丈介　小高 知宏（福井大学大学院工学研究科）
諏訪 いずみ（仁愛女子短期大学）　白井 治彦（福井大学工学部）

1　はじめに

本研究の目的は Word2Vec の CBoW モデルから獲得した単語ベクトルを用いて，著者の分からない小説の著者を推定することである．これまでの研究では，作品中に使用されている語彙に対応する単語ベクトルから，作品と著者の類似度としてマハラノビス距離を求め，著者推定を行った．得られた結果として，検証データにおいても著者 A と著者 A の作品 α 間の類似度は小さくなり，その他の著者と作品 α 間の類似度は大きくなることが分かった [1]．しかし，このコーパス中には評論や手紙などといった小説作品以外の作品も含んでいた．そこで，対象とする作品をコーパスのうち小説であると判断した 917 作品にコーパスの特性を限定し，著者推定に与える影響を明らかにすることを本研究の目的とする．

2　単語ベクトルを用いた著者の特徴分析方法

コーパスを Word2Vec の CBoW モデルで学習することで，その中間層では多次元の単語ベクトルを獲得できる．本研究では，著者 A の作品 α の特徴量 $\boldsymbol{f}^{(A,\alpha)}$ を以下で定義する．

$$\boldsymbol{f}^{(A,\alpha)} = \frac{1}{|X^{(A,\alpha)}|} \sum_{i \in X^{(A,\alpha)}} \boldsymbol{x}_i^{(A,\alpha)} \tag{1}$$

ここで $X^{(A,\alpha)}$ は著者 A の作品 α に含まれる語彙の重複を許した集合，$\boldsymbol{x}_i^{(A,\alpha)}$ は i 番目の語彙の単語ベクトルである．ある作品 β と著者 A との類似度を，以下に与えるマハラノビス距離 $d_A^{(\beta)}$ で定義する．

$$d_A^{(\beta)} = (\boldsymbol{f}^{(\beta)} - \boldsymbol{\mu}_A)^T \Sigma_A{}^{-1} (\boldsymbol{f}^{(\beta)} - \boldsymbol{\mu}_A) \tag{2}$$

$\boldsymbol{\mu}_A$ は，著者 A の学習データから算出した単語ベクトルの平均，Σ_A は単語ベクトルの共分散である．

3　著者推定法

まず，著者 A の特徴量 μ_A に対する全著者の学習データの特徴量のマハラノビス距離 d_A を求めた．この距離データでは，明らかに著者 A の作品が小さい距離にあることが確認できる（図 1 参照．図 1 では芥川）．よって，A 以外の著者の作品が誤認識されない距離を閾値 θ_A とする．この手順を各著者に対して行い，著者毎の閾値を決める．そして，各著者の特徴量に対する全検証データのマハラノビス距離に閾値 θ を設定し，著者 A の閾値 θ_A を下回る作品の著者は A であると推定する．複数の著者の閾値を下回った場合は，最小の距離となる著者であるとする．

4　著者推定実験

対象とする著者の，青空文庫で公開されている全ての作品から，句読点，記号やルビ，出版年月日などの情報を削除し，フリーの形態素解析器 MeCab でわかちがきした．わかちがきした文章を，Python ライブラリである Gensim の CBoW モデルを利用し，埋め込み次元数 $N = 100$，窓数 $w = 5$ で 300 回学習を行った．CBoW の学習では，小説以外の作品も含む全ての作品を用いて学習を行った．

著者推定に用いるコーパスには 10 著者の小説作品 917 作品を用いた．このうち，芥川龍之介，坂口安吾，太宰治，牧野信一，宮沢賢治の 5 名は作品数が 101 作品を越えている．この 5 著者の作品のうち，101 作品を学習データとし，残りを検証データとした．ただし，学習データと検証データの作品の単語数に偏りがないようにした．他の 5 著者の作品は全て検証データとした．

5　考察

学習作品のマハラノビス距離のヒストグラム（図 1）において，実際の著者の作品は距離が近くなる傾向が明確に見られた．このことより，単語ベクトルから求めた著者の特徴に対する作品の特徴のマハラノビス距離は，著者の特徴を反映しているといえる．推定結果は当日発表する．

参考文献

[1] CBoW から見た近代小説文章における著者の特徴と著者推定，竹中志織，黒岩丈介，小高知宏，諏訪いずみ，白井治彦，2020 年度電気・情報関係学会北陸支部連合大会，F-2-4-12

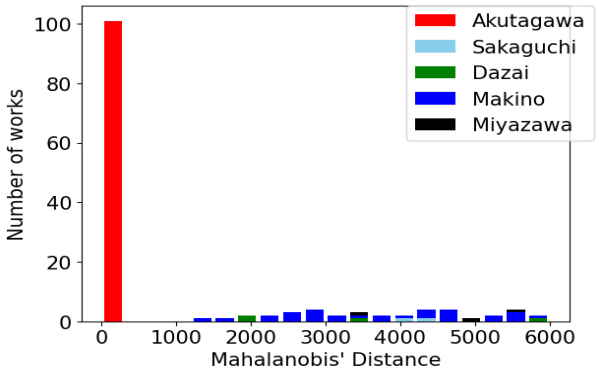

図 1: 芥川の特徴量に対する 5 著者の学習データのマハラノビス距離