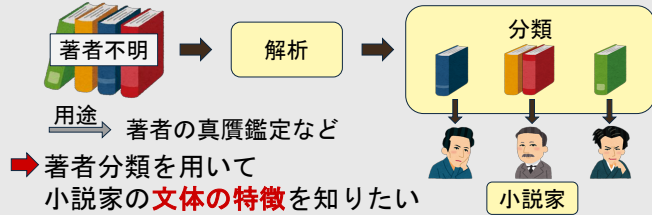


形態素解析を用いた著者分類

明治大学 総合数理学部 現象数理学科 池田研究室

1. 背景

著者分類とは 著者不明の文書を分類し、著者を推測



先行研究: n-gramを用いた著者の判別

[松浦司, 金田康正, "近代日本文学作家8人による文章のn-gram分布を用いた著者判別", 2000.6.1]

隣接したn個の文字 (n-gram) の出現確率分布を指標とする
例) 3-gram
吾輩は猫である。 → 吾輩は 猫は猫 は猫で 猫であ ある。 ある。
ある日の暮方の事である。 → ある日 ある日 の暮 の暮方 暮方の 方の事 方
の事 事である である ある。 ある。
→ それぞれの3-gramの出現確率を比較し、同じ著者によるものか判別する
データ: 明治から昭和初期にかけての日本近代作家の作品計92作品
指標: 1-gramから10-gram分布
結果: 指標に2-gram, 3-gram分布を用いているとき著者判別の精度が高い
最良の場合では最短2万文字のテキストで100%の精度

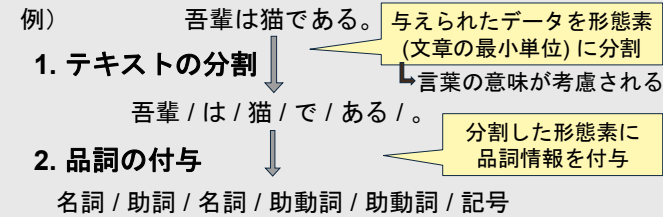
問題点
• n-gramの区切りに言葉の意味は考慮されていない
• 著者の特徴が文字単位に細分化されており複雑

○判別精度: 高 ✕判別基準が理解しづらい

➡改善策 言葉の意味を考慮した指標
人間が解釈しやすい指標が必要

2. 形態素解析—自然言語処理の一過程

自然言語処理...普段使っている言葉 (日本語、英語等) をコンピュータが理解できるようにする処理



形態素解析によって得られた指標

- 文末表現 (使用率上位10個)
- 文頭表現 (使用率上位10個)
- 文頭表現が主語である割合
- 文頭表現 & 主語の割合
- 2・3文字の形態素 (各使用率上位10個)
- セリフの割合
- 各品詞の割合
- 間の取り方
- 各文字種の割合 (ひらがな・漢字・記号等)
- 文字数

言葉の意味を考慮した指標

人間が解釈しやすい指標

指標ごとに作品をクラスタリングで分類

➡著者の判別 & 文体の分析を行う

今回はMecabという解析器を使用し、辞書にはIPA辞書を用いた。IPA辞書は基本的な日本語の文法、専門用語などの固有な名詞に強い。

3. 目的

明治～昭和初期の小説家10人の各20作品において
形態素解析で得られた指標を用いて

- 著者分類が可能であるのか検証
- 各著者の文体の傾向分析

データは青空文庫より新字新仮名遣いに変換済みのものを使用し、執筆時期が近く、作品数が豊富で十分な量のデータが確保できる小説家を選択した。したがって今回の実験において、時代による語彙や言葉遣いの変化は考慮されないものとする。

4. 結果—作者ごとのクラス割合

※一部指標の結果は省略

以下の2つの指標では一部著者分類が可能!

□各文字種の割合

□...0.25以上0.5未満 □...0.5以上0.75未満 □...0.75以上

クラス	芥川	泉	海野	岡崎	菊池	佐々木	島崎	太宰	夏目	森
1	0.3	0.15	0.05	0.0	0.1	0.0	0.0	0.1	0.2	0.15
2	0.25	0.05	0.05	0.0	0.2	0.0	0.5	0.05	0.2	0.3
3	0.25	0.15	0.1	0.0	0.15	0.0	0.0	0.0	0.05	0.05
4	0.15	0.0	0.2	0.3	0.35	0.0	0.35	0.15	0.45	0.2
5	0.05	0.6	0.0	0.0	0.15	0.0	0.0	0.0	0.0	0.0
6	0.0	0.05	0.35	0.05	0.0	0.0	0.0	0.25	0.0	0.0
7	0.0	0.0	0.25	0.05	0.0	0.3	0.0	0.25	0.0	0.05
8	0.0	0.0	0.0	0.55	0.0	0.0	0.15	0.2	0.1	0.1
9	0.0	0.0	0.0	0.05	0.05	0.0	0.0	0.0	0.0	0.15
10	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.0

➢泉・岡崎・佐々木・島崎は文体が特徴的である

- 1つのクラスに0.5以上の作品が分類されている
- 特にクラス10には佐々木の作品しか分類されていない
- ひらがなが多く、漢字・記号が少ない

分類可

□2・3文字の形態素

クラス	芥川	泉	海野	岡崎	菊池	佐々木	島崎	太宰	夏目	森
1	0.85	0.3	0.4	0.55	0.55	0.0	0.2	0.55	0.1	0.5
2	0.1	0.5	0.2	0.25	0.4	0.9	0.2	0.35	0.25	0.25
3	0.05	0.15	0.2	0.1	0.0	0.1	0.25	0.1	0.05	0.1
4	0.0	0.05	0.0	0.0	0.0	0.0	0.0	0.0	0.05	0.0
5	0.0	0.0	0.15	0.0	0.0	0.0	0.1	0.0	0.15	0.05
6	0.0	0.0	0.05	0.05	0.0	0.0	0.05	0.0	0.1	0.05
7	0.0	0.0	0.0	0.05	0.0	0.0	0.05	0.0	0.05	0.05
8	0.0	0.0	0.0	0.0	0.05	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.2	0.0
10	0.0	0.0	0.0	0.0	0.0	0.0	0.05	0.0	0.0	0.05

➢芥川・佐々木は文体が特徴的である

- 1つのクラスに0.75以上の作品が分類されている
- 芥川: 1作品の文章量が少なく、使用される語彙が限られている
- 佐々木: 「まする」「ござり」「ござん」といった話し言葉を多用している

➢島崎・夏目は文体の特徴がない

- クラス割合の最高値が0.25でありクラス分類の偏りが小さい
- 両者ともに1作品の文章量が多く、使用される語が幅広い

分類可

佐々木に関する2つの指標の結果には

ひらがなや話し言葉が多く、漢字や記号が少ない
ことが影響していると考えられる

実際に、佐々木の作品は芸術性に重きを置いた純文学ではなく、娯楽性を重視した大衆向け小説である。

5. 結論

1. 特定の小説家において著者分類は可能

→著者ごとに特徴となる指標が異なる

2. 佐々木の文体を分析できた

➡形態素解析とクラスタリングによる
著者分類は有用である