

人工知能の文章に文体はあるか

——チャットAIが生成した物語文章の文体分析——

○井関 龍太
(大正大学)



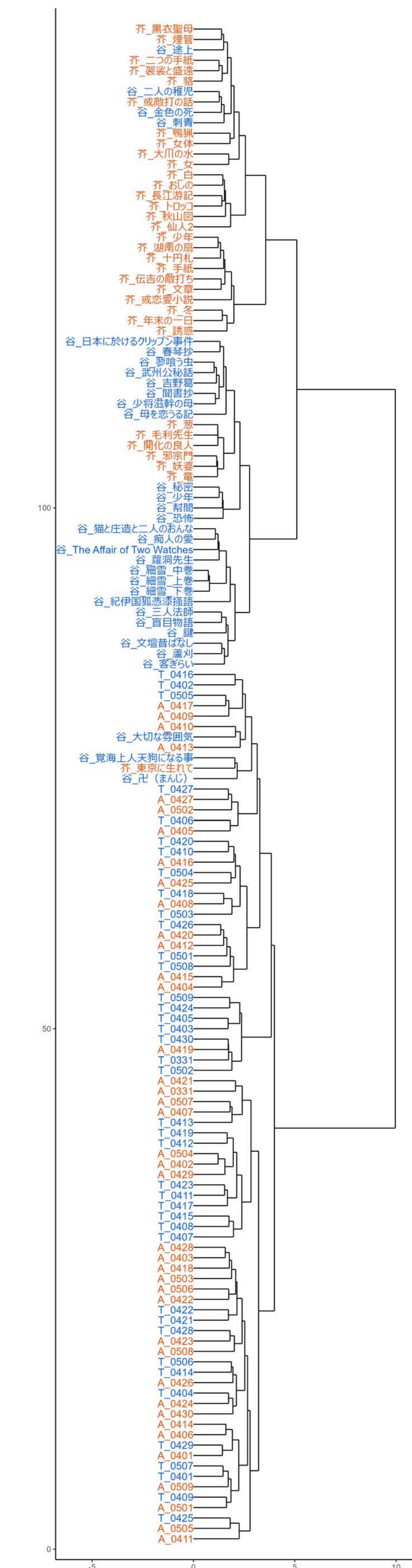
問 題

- ▶大規模言語モデル:大量のテキストデータの学習に基づく自然言語処理モデル(ChatGPT, Bard, Bing AIなど)
 - ・非常に流暢な文章を生成できる
 - ・(失敗することもあるが)まとまりのある文章に見える→**一個人による文章**と同様のまとまりがある?
特徴的な書きかたのパターン(文体)を持っているか
- ▶本研究の目的:大規模言語モデルが人間の書き手と比較可能な特徴を備えた文章を生成するのか検討する
 - ・著者識別でよく用いられる助詞の使用を分析(金, 2002)
 - ・**芥川龍之介風**と**谷崎潤一郎風**の文章を生成
 - ・実際の**芥川龍之介**と**谷崎潤一郎**の文章と区別されるのか検討a)まとまった文体を持たない
→AIによるテキストはあちこちに分散する
b)まとまりがある+特定の作家を模倣
→模倣した作家とまとまる
c)まとまりがある+独自の文体
→AIのみでまとまる

方 法

- 材料:Bing AIとのチャットによって生成した物語文章
 - ・**芥川龍之介風**と**谷崎潤一郎風**の各40の物語
 - ・プロンプト:「～風のショートストーリーを作れますか?」
ただし, Bing AIが物語を作らないこともあるので, 適宜追加で指定したり文言を変更したりした
 - ・比較用の文章:青空文庫に収録された**芥川龍之介**と**谷崎潤一郎**が著者である33の小説作品
(アクセス時点で芥川作品は132件, 谷崎作品は33件だったので, 前者からはランダムに33件抽出)
- 分析:**助詞**の使用頻度についてクラスター分析
 - ・各テキストについて形態素解析:MeCab 0.996 (工藤, 2006)とNEologd 2020-8-20(佐藤, 2020)を使用
 - ・テキストごとに助詞の使用数を計上:テキストの長さが異なるため(AIは平均40.15文, 作家は平均250.20文), 助詞の出現頻度を1,000形態素あたりの出現数に換算
 - ・助詞の出現数を平方Jensen-Shannonダイバージェンスによって距離行列に変換
 - ・ワード法によるクラスター分析

結果と考察



人間の作家による
テキスト

Bing AIによる
テキスト

- ▶作家によるテキストとAIによるテキストで大きくクラスターが分かれた
 - ・作家による4作品だけがAIのクラスターに含まれた
 - ・作家が大半を占めるクラスターにはAIの作品なし→作家とAIでは**文体(助詞の使用の傾向)**が異なる
解釈1:作家と区別できるような独自の文体がある?
解釈2:特徴のない文体がまとまった(文体がない)?
- ▶作家によるテキストの中では, **芥川龍之介**と**谷崎潤一郎**でおよそクラスターが分かれた
 - ・**芥川クラスター**:26/30が芥川作品
 - ・**谷崎クラスター**:26/32が谷崎作品→それぞれの作家に特徴的な文体がある
- ▶AIによるテキストの中では, **芥川風**と**谷崎風**は明確には分かれなかった
 - ・**芥川風**・**谷崎風**でまとまりができているところはあるが, 大きなクラスターを作らない(AI作品だけを対象に再分析しても同様)→**芥川風**と**谷崎風**で文体を書き分けてはいない
- ・本研究では, あくまで「～風」の物語を作る指示を与えた
- ・文体をまねるようAIを調整すれば書き分けも不可能でないかも