

テキストマイニングを用いた日本文学作品 の特徴量分析による難易度判定

総合情報学研究科 博士前期課程 1年

王梓丞

研究背景

多くの日本語を学ぶ学習生は、日本の文学作品を読んで見たいと思っている。

しかし、学習している言語で文学作品を読むことには困難が伴う。

その理由は、文学作品にはでてこない幅広い語彙と会話などの生きた言語表現が出てくるからである。

研究目的

テキストマイニングによる分析を通して、日本文学作品が日本語学習者に適しているかどうかを検討する。

日本語の学習者を対象とした日本の文学作品をテキストマイニングにより分析する。学習者の読解難易度を推定し、日本語の学習者に適する文学作品を選ぶ。

先行研究

河原井、平本、蓬萊 「中世日本文学に対する自然言語処理」、FIT2018（第17回情報科学技術フォーラム）

自然言語処理技術で日本文学の分野に関する分析の可能性を探究

花畑、青野 「語彙と文脈に着目した文学作品の著者推定」、言語処理学会年次大会2019
文学作品の著者の語彙と文脈の特徴に着目した著者推定手法を提案

夏目漱石の作品をモデルにした理由

1. 夏目漱石は日本を代表する作家であり、日本の学校の教科書にも作品が載っている。
2. 夏目漱石の作品は外国でも広く知られている。

仮説

上記から、日本語を外国語として学ぶための教科書として適しているのではないか。

日本語能力試験（JLPT）

日本国際教育支援協会が運営

日本語を母語としない人の日本語能力（文字・語彙・文法など）を測定し認定する試験である。

日本語能力試験のレベル(n1～n5：上級～初級)によって、必要となる漢字・語彙の数が異なる。

➤本研究では、JLPTの主要な単語と文法を判定し基準とする。

テキストマイニング

単語の出現頻度に基づいて、N1とN2の単語の出現頻度と分布の状況を考察する。

単語間の関係を考察し、文法の構造とその組み合わせを加え、作品における文法の難度を考察する。

一般的に、文章の長さの平均的な長さを調べてみると、文章の長さが長いほどわかりにくいと考えられる。

一文につき文章に出す平均値を考察する。一般的に、文章は長ければ長いほど読みにくく、読みやすい文章ほどスッキリと短くまとまっている傾向があると思われる。

提案手法

1. 夏目漱石の作品を中心にして、アンケートを通して、日本語学習者の読解能力に適した難易度の作品を推定する。
2. 国際日本語能力試験N3からN1までの語彙及び文法を参考にして、アンケートから選んでいる作品を分析し、教師データを作る。
3. 教師データにより、他の前処理された文学作品(夏目漱石以外の作品)は、N1、N2、N3 の3つのレベルで判断できる新しいアルゴリズムの提案手法を考察する。

アンケート例

名前	性別	学歴	職業	専攻	日本能力試験成績	日本語学習時長
赤崎	男	大学	学生	経済学	N2	2年
夏目漱石作品			読むこと	読む難易度 (* 1)	難易度理由 (* 2)	作品評価
『吾輩は猫である』			0	2	2	
『坊っちゃん』						
『草枕』			0	5	1 3	難しい
『二百十日』						
『野分』						
『虞美人草』			0	5	1 3	読み始めは難しかった
『坑夫』						
『三四郎』						
『それから』						
『門』						
『彼岸過迄』						
『行人』						
『こゝろ』						
『道草』						
『明暗』						
			(ありの場合○)	(1から5に入力)		
注		(* 1) 1 : 簡単、2 : やや簡単、3 : 微妙、4 : やや難しい、5 : 難しい (* 2) 選択肢 1 : 単語難し 2 : 単語簡単 3 : 文法難しい 4 : 文法簡単 5 : 段落読む難い 6 : 段落読 7 : 文章意味理解 8 : その他				

アンケートによる難易度の分類

作品	簡単	やや簡単	普通	やや難しい	難しい	難易度決定
吾輩は猫である	50%	25%	25%	0%	0%	簡単
坊っちゃん	0%	25%	50%	25%	0%	普通
草枕	0%	0%	0%	25%	75%	難しい
虞美人草	0%	0%	0%	0%	100%	難しい
注：今まで4人アンケート例になります						

学習データの設計

作品	n1単語 比率	n2単語 比率	n1文法 比率	n2単語 比率	...	句長	難易度 決定
吾輩は猫 である							簡単
坊っちゃん							普通
草枕							難しい
虞美人草							難しい
...							

入力と出力、構築される学習モデルの関係

入力変数 $\ddot{x}$

N2の単語比率

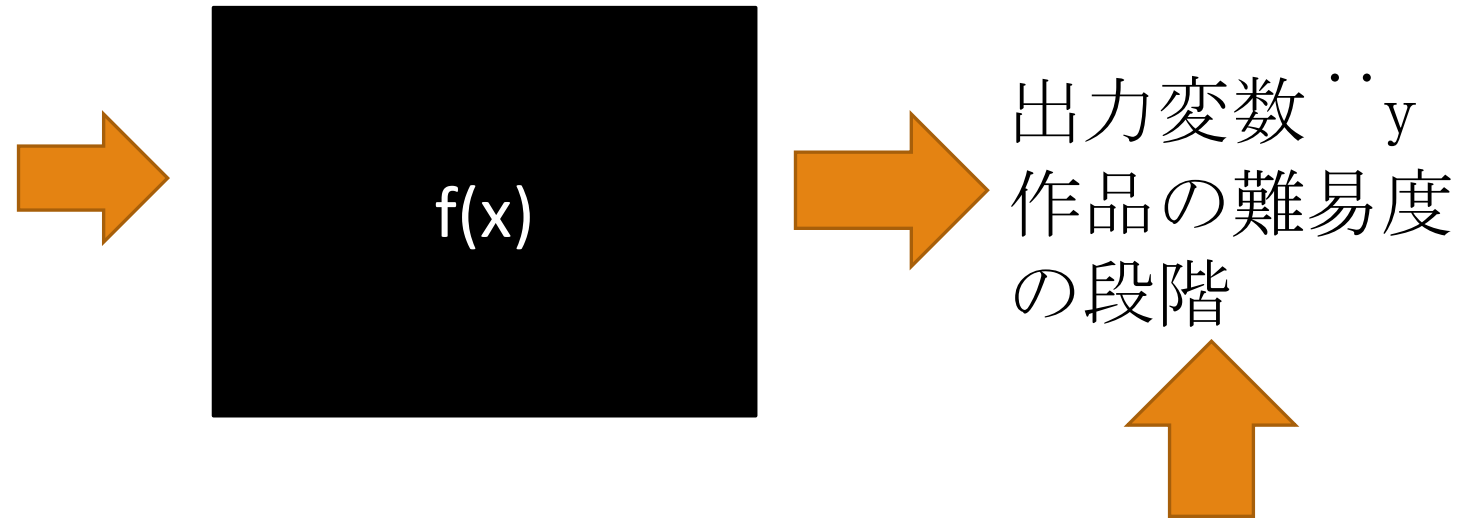
N1の単語比率

N2の文法比率

N1の文法比率

句長

...



出力変数 $\ddot{y}$
作品の難易度
の段階

教師変数 $\ddot{t}$
アンケートによる作
品難易度の段階

予定

JLPTのN1コアワード700単語、N2コアワード700単語、N3コアワード600単語、N1コア構文構造60文、N2コア構文構造90文、N3コア構文50文などを利用する。

夏目漱石の作品は、アンケートによって、簡単、やや簡単、普通、やや難しい及び難しい五段階に分ける。

夏目漱石の作品から単語難度分布や文法難度分布などを求める。

参考文献

1. Yamada M, Murai Y. Story visualization of literary works[J]. Journal of visualization, 2009, 12(2): 181-188.
2. Argamon, S., Saric, M., and Stein, S. (2003). Style Mining of Electronic Messages for Multiple Authorship Discrimination: First Results. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 475-480.
3. 宮内拓也, 浅原正幸「日本語における名詞句の情報構造と語順の相関についての統計的検討」『自然言語処理 = Journal of natural language processing』言語処理学会 編 27(2):2020.6 p.361-381
4. 名畑目真吾「外国語教育に関する研究動向 : 自然言語処理研究との接点に着目して」『筑波教育学研究』筑波大学教育学会 編 (17):2019.3 p.33-52
5. 柳井孝介、庄司美沙『Pythonで動かして学ぶ自然言語処理入門』翔泳社、2019.1
6. 杉浦清人「文学研究におけるデジタル・ヒューマにてイーズの可能性: 文章心理学・計量文献学・マクロ分析」、『れにくさ: 現代文芸論研究室論集7』、2017、 p. 80-96
7. 田畑智司「FLOB コーパスの意味構造: 確率論的トピックモデルによる言語使用域の特徴付け」言語文化共同研究プロジェクト、 2017、 p. 5-21
8. 木村美「計量文体論の手法を用いた文芸テキスト分類—James Tiptree, Jr. と Ernest Hemingway—」『文学研究論集47』, 2017、 p. 13-31

ご静聴ありがとうございました。