5P-8

注釈の自動生成による青空文庫の読書支援

速水　秀平†　　井上　潮‡
東京電機大学工学研究科情報通信工学専攻†
東京電機大学工学部情報通信工学科‡

1　はじめに

著作権の消滅した近現代文学作品が読めるサイトとして青空文庫がある.近年ではパソコン以外にも,タブレット,スマートフォン,電子書籍リーダなどの電子デバイスが普及し,青空文庫,ひいては近現代文学作品を読む為の敷居が大幅に下がったと言える.
しかし,50 年以上前に書かれた小説には現代では使われる事の少ない語句,固有名詞が数多く存在し,大半の小説は,読書経験の浅い者には読む事が難しい.そこで,予めその語句の現代においての利用度を判別し,一般的ではない語句には「注釈」という形で,読者の読書ペースを乱すことなく文章の理解を支援するシステムを考案した.

本研究の最終的な目標は読書支援システムを構築する事であるが,現時点では語句ごとの難読性を判定する手法の検討を行っている段階である.

2　関連研究

近現代文学作品の難読性を研究した物として,河内による[1][2]「羅生門」と「走れメロス」の難読語の個数による教材価値の判定を行った物がある.これによると,この 2 作品は教材として用いられやすい事から現代語の割合は多い物の,やはり現代では滅多に用いられない品詞も一定数存在している事が分かるという結果になっている.

3　提案手法

まず、小説中で使われている語句に対し,その語句が現代で使われている頻度を調べる,この時,使われている頻度が非常に低い場合,一般的に目に触れる機会がほとんど無いという事である為,語句の意味を知っている者も少ないという仮説を立てた.この手法がどれだけ妥当であるかを判定する為に,現代語コーパスによる語句のヒット率を測定する.

Reading support of Aozora Bunko by automatic annotation generation.
† Shuhei Hayami Tokyo Denki University
‡ Ushio Inoue Tokyo Denki University

まず,青空文庫にある小説に対し,形態素解析を用いて品詞ごとに分割する.

3.1 形態素解析

小説内の語句の抽出には形態素解析を用いる.今回,形態素解析用の辞書としては近代文語 UniDic(ver.1.3)を用いた.
理由としては青空文庫と太陽コーパスを基に作成された辞書であり,本研究の用途に非常に適している為である.また,形態素解析ツールとしては辞書の使用に推奨されている Mecab (ver.0.996)を用いる.
本来であれば精度の検証を行う必要があるが,現段階ではそれほど高い精度が求められていない事と,辞書の作成方法や,サイト上にある精度の検証結果からある程度の信頼性がある事から[3],見送った.

3.2 コーパス

コーパスには,日本語版 Wikipedia の 2013 年 11 月 7 日時点での全記事データ(約 88 万記事)と,Baidu コーパスの 2008 年 1 月から 2010 年 7 月までの物を用いた.これは,Baidu のクローラが集めたブログおよび掲示板の文章を形態素解析で分割し,語句ごとに出現回数をまとめたものである.使われた形態素解析ツールは Mecab(var.0.98),辞書は IPA 辞書(ver.2.7.0)である。ここで,Wikipedia は百科事典であり,Baidu コーパスはインターネット上のブログや掲示板で使われた文章を集めた物である.

4　測定手法

今回は森鴎外の「舞姫」の全文を形態素に分け,それらの中から名詞に該当する物をコーパス内で検索し,その出現回数を調べた.また,品詞に分解した物をそのまま検索すると旧字体の物はヒット率が下がるため,その語句の語彙素で検索を行った.

5　考察

対象となる小説の形態素解析を行った結果,名

詞に該当する物は 1033 語であった.これをWikipedia コーパスで判定した所,最も出現率が高かったものが 11,807,227 回の「年」であった.

また,出現回数が 0 であった品詞は 36 個あった.表 1 は出現回数が 0 件であった物のうち,小説内で出現した順に 10 件を抽出したものである.

Baidu コーパスで小説内の名詞を検索した所,0 件以上の物は 771 件であり,最もヒット数が多かったものは 486,931 回の「人」であった.

この内,Baidu コーパス内で 1 件以上である品詞の中で Wikipedia コーパス内でのヒット数が最も少なかった物から 15 件を抽出した物が表 2 である.

表 1 を見ると,速水自身の知識ではあるが,意味を知っている語句は 1 件も無い為,少なくともヒット率 0 件の物は注釈が必要であるとみなせる可能性が高い.

また,この内「欷歔」は「欷」が常用漢字では無い.この事から,ヒット数が少なく,常用漢字以外の漢字を含んだ品詞は難読性の高い語句としてみなせる可能性がある.

ただし,ニルアドミラリーという品詞は,このまま検索してもヒットしないものの,「ニル・アドミラリ」では 4 件ヒットした.この事から,同義語の検索も今後手法として取り入れる必要があると考えられる.

トヨタロウという品詞は小説中に登場する人物名であり,今回の研究において除外すべき品詞である.この事から,例えば小説内での出現頻度が高い物の,コーパス内での出現率が極端に低い物は小説固有の名詞であるとみなす等,例外を考える必要がある.

表 2.を見ると,Baidu コーパス内で 1 件以上ヒットした語句のうち,Wikipedia コーパス内での最低ヒット率は 45 件である.つまり 44 件以下の語句の Baidu コーパスでのヒット率は 0 件という事になる為,2 つのコーパスを用いた測定には効果があるとみなせる.

また,「百科事典」と「ブログと掲示板」という文章の利用目的の違いによるヒット率の違いも見られる.例えば上靴という品詞は Wikipedia では 52 件なのに対し,Baidu コーパスでは 4 件と差が大きい.この事から,現代の文章であるものの,違う用途で使われる文章内でそれぞれ検索をする事で,難読性の判定の精度を高められる可能性がある事が分かる.

以上の事から,今後の研究方針としては用途の違う複数の日本語コーパスから検索すると共に,同義語の検索や、登場人物名などを除外するための判定の手法の考察を行う.

表 1. Wikipedia コーパスに含まれなかった名詞(36 件中の 10 件)

単語
物学び
ニルアドミラリー
惨痛
トヨタロウ
鈴索
レーベマン
獣苑
モンビシュー
欷歔
獣綿

表 2.Wikipedia コーパスと Baidu コーパスの比較(Wikipedia コーパス内でのヒット率の低い物から 15 件)

単語	Wikipedia	Baidu
早足	45	54
悪阻	49	26
上靴	52	4
身の程	54	51
会釈	91	32
気色	97	195
一抹	98	68
貸家	102	20
御身	112	6
一瞥	113	15
見覚え	144	312
面持ち	145	44
酒家	159	10
放言	161	17
独り身	162	53

参考文献

[1] 河内 昭浩(2012)「教材「羅生門」の語彙研究」,安田女子大学紀要 p40,193-202,2012-02-28
[2] 河内 昭浩(2012)「教材「走れメロス」の語彙研究」, 安田女子大学紀要 p41,189-198,2013-02-28
[3] 小木曽智信(2008)「近代文語文を対象とした形態素解析辞書・近代文語 UniDic」p29