様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月 ２２ 日現在

機関番号：５２１０１
研究種目：挑戦的研究（萌芽）
研究期間：2017～2020
課題番号：１７Ｋ１８４９９

研究課題名（和文）テキストマイニングの手法を用いた説話集の性質分析と分類

研究課題名（英文）The classification of narrative literature using the technique of the text mining

研究代表者
平本　留理（Hiramoto, Ruri）

茨城工業高等専門学校・国際創造工学科・准教授

研究者番号：２０３４２４６２

交付決定額（研究期間全体）：（直接経費）　　2,500,000円

研究成果の概要（和文）：９つの説話集を対象に、個々の本文データベースを作成して形態素解析を行い、それぞれの作品における特徴的な語彙を抽出した。『古今著聞集』においては、抄入とされている説話の検証を機械的に行い、従来の指摘を裏付ける結果を得た。さらに、三大説話集を対象に、階層的クラスタリングと非階層的クラスタリングの２種類の手法を使って、それぞれの説話集の性質分析と、説話集間の類似性の検証を行った。最終的には、９作品を対象に、説話集同士の類似性をもとにした分類を試みた。その結果、文体的類似性のみならず、内容的類似性の観点からも、ある程度の分類結果を導き出すことができた。

研究成果の学術的意義や社会的意義
本研究の意義は、既存の文学研究の手法とは違った観点から、文学作品の性質に迫ろうとしたところにある。情報工学の手法である「形態素解析」と「テキストマイニング」を用いることにより、これまでの文学研究では難しかった複数の作品を同時に検証することを可能にしている。膨大な情報量の中から人間の目では追い難い特徴を見出すことや、機械的な分類結果がこれまでの諸氏の先行論をどこまで裏付けられるのかの検証を試みたのだが、今後の文学研究の一手法として、ある程度の可能性は見出せたものと考える。

研究成果の概要（英文）：I made nine text databases of the narrative literature. And I performed morphological analysis of them. I found out characteristic vocabulary from each text. With the Ward method and X-means clustering, they were classified into several groups.
There are some insertion stories in “Kokonchomonju”. Using SVM, I inspected whether those stories were the insertion. As a result, I was able to support conventional hypotheses.

研究分野：中世説話文学

キーワード：説話集　本文データベース　自然言語処理　テキストマイニング

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

筆者はこれまで『十訓抄』や『古今著聞集』の説話文学研究を手掛け、「説話」という分野の持つ性質を明らかにしようと試みてきた。その中で、早くから各作品における語彙の使用傾向の特徴を探る手法を取り入れてきた。しかし、当時は本文のデータベース化が今ほど進んでおらず、自然言語処理分野に関する知識もなかったため、該当する語を抽出するのに、索引などを利用してすべて手作業で行っていた。この方法では時間的にも限界があるため、一部の語彙の傾向を探るに留まらざるを得ず、思うような成果を挙げることができなかった。

そのような時に、情報工学の自然言語処理の分野において、まさに筆者が必要としているような技術を用いた研究があることを知り、これを文学作品研究の分野に応用できないかと考えるようになった。

特に説話文学の領域においては、さまざまなタイプの説話集がある中で、分野全体を俯瞰して類似性や影響関係を探る研究が十分に行われているとは言い難い。これは先に述べたように、既存の文学研究の手法では研究者が追える範囲に限りがあり、一度に多くの文学作品を検証の対象とすることが不可能であったことに起因すると考えられる。

情報処理分野の技術は、ビッグデータの扱いを可能にする。特に、使用語彙の傾向を、数量的にも視覚的にもよりわかりやすく示すことができるテキストマイニングの手法は、文学研究を手掛ける者にさまざまな「気付き」を与える助けになるのではないかと考えた。本研究はそのような着想のもと始めたものである。

２．研究の目的

本研究は、説話文学研究の領域において、個々の説話集の性質や、説話集相互の類似性を明らかにするための手段として、「テキストマイニング」の手法を取り入れようとするものである。説話集はさまざまな文献から説話を収録していることが多く、その作品研究において他文献との影響関係を探ることが欠かせないが、現状の分類として一般的なものである「仏教説話」か「世俗説話」かの分類にはこの観点がない。これは、分野全体を俯瞰し、個々の説話集の位置づけをはかる観点として不十分だと考え、これらの新たな分類を提唱することを最終目的とした。

３．研究の方法

まずは分類の対象となる説話集の本文データベースを作成する。その本文データの形態素解析を行い、結果として得られた個々の説話集の使用語彙傾向をもとに、複数のテキストマイニングの手法を試す。説話集の性質や説話集間の類似性をあぶりだすのに適したテキストマイニングの手法を活用し、対象とした複数の説話集をいくつかのグループに分類する。

４．研究成果

（１）本文データベースの作成

期間中、９つの説話集の本文データベースを完成させた。これは主に岩波大系本をテキストとしたものであるが、国文学研究資料館などで公開されているものとは違い、その後形態素解析を行うことを前提に、それに適した形に改めたものである。

最初に説話集 13 作品の本文の一部を用いて形態素解析を試行するためのサンプル文を用意し、実際に解析を行ったうえで、解析上の問題点を抽出し、どの程度本文に手を加えれば解析を行った際にエラーが少なくなるのかを検討しながらデータベースを完成させた。ただし、当初の予定では 13 作品すべての本文データベースを作成する予定であったが、時間と予算上の都合から完成させることができたのは 9 作品に留まった。

（２）形態素解析による特徴語の抽出

各説話集の本文において、それを構成する単語ごとに切り分ける形態素解析を行った。それぞれの説話集の性質を浮き彫りにするべく、まずは巻や篇などを単位として、そこで使用されている特徴的な語彙を抽出した。これらの分析により、各説話集のそれぞれの巻がどのような内容的特徴を持つか、また各説話集の文体的特徴がどのようなものか、その傾向の概要をつかむことができた。また、『今昔物語集』など一部の説話集においては、巻と巻との間の類似性を確認することもできた。

ここでの分析結果は、これまでの先行論から大きく外れることはなく、おおよそそれらを裏付

ける結果になったと言えよう。ただし、一部では、やはり思いがけない特異性や類似性が浮かび上がってきたところがあり、先行論に指摘の見られない部分についてさらなる検証を行う必要性が出てきた。これらは研究者の「気付き」を促した要素となり得るものであり、この結果をもとに次なる研究へとつなげる可能性を見出した。

（３）『古今著聞集』における抄入部の検証

テキストマイニングの手法を説話集の性質分析に活かす試みの一つとして、『古今著聞集』における抄入部の検証を行った。『古今著聞集』にはもともとの編者橘成季の手ではなく、後に他者の手によって説話を挿入されたと思われる抄入部がかなりの数あることが知られている。その中で、おそらく抄入であろうとの指摘はありながらも、その出典がはっきりせず、疑義のまま考察の進んでいない４つの説話が本当に抄入か否かを、コンピュータに判定させるということを試みた。これは SVM という手法を使ったもので、疑義のある４説話を除き、オリジナルと思われる説話グループと、抄入と思われる説話グループの特徴をそれぞれ機械に学習させ、最後に４説話の一話一話について、それがオリジナルグループと抄入グループのどちらに入るものなのかを機械的に判定させるというものである。

結果、コンピュータは従来の指摘どおり、４話いずれも「抄入」であると判定した。ここで興味深い点は、これまでの指摘がなされてきた背景には、説話の配列方法などの問題が主に含まれていたのに対し、本研究で機械が導きだした過程にはその点が考慮されていないという点にある。つまり、既存の考察過程とは全く別のアプローチにおいて、この４話がオリジナルとは違うと見なされる要素があったために「抄入」の判定がなされたということになる。その判定要素が何であったのかは現在検証中であるが、これは、このような研究手法が新たな文学研究のあり方の一つとして提言し得るものであることを示す検証結果であったと考えている。

（４）三大説話集における性質分析

全９作品での分類を行う前に、『今昔物語集』『宇治拾遺物語』『古今著聞集』の三大説話集における作品の性質分析と、それぞれの説話集間の類似性に関する検証を行った。類似性に関しては、性質の近いもの同士をグループ化していくクラスタリングの手法によって示すことにした。

まず各説話集のテキストを巻や篇という単位で分け、階層的クラスタリングと非階層的クラスタリングのそれぞれにおいてどのような分類がなされるかを検証した。その結果、同一説話集内の巻同士は同じクラスタ内に入ることが多く、作品全体としての文体・内容面での類似・統一が見られることを確認した。しかしながら、複数の編者が関わったと思われ、また巻ごとのテーマ性がはっきり分かれている『今昔物語集』は、同一説話集の中でもさらに傾向が大きく２ないし３に分かれることがわかった。さらに、３つの説話集間における性質の類似性で見ると、『今昔物語集』と『宇治拾遺物語』は完全に離れ、その間に『古今著聞集』が位置するということが確認できた。ただし、『古今著聞集』はこの２作品の中間的存在というわけではなく、性質上はかなり『宇治拾遺物語』寄りであるという結果になった。これは各説話集に収録される個々の説話を単位とした分析でもほぼ同様の結果であった。さらに、品詞ごとの分類にすると、より文体的な面での類似性か内容的な面での類似性かがはっきりすることも確認できたうえ、傾向が出づらい品詞もはっきりしてきた。傾向の出にくい品詞の単語について、これらを除いた検証を行なえば、より明確に説話集間の類似性を浮き彫りにすることができる可能性が考えられる。

（５）説話集９作品の分類

先に述べた三大説話集に加え、『日本霊異記』『江談抄』『富家語』『中外抄』『十訓抄』『沙石集』の６作品を加えた９作品での分類を試みた。最終的に、階層的クラスタリングにおいては Ward 法による分類を、非階層的クラスタリングにおいては x-means による分類を行った。これにより、それぞれの説話集間の類似性の強さを、視覚的にもわかりやすく距離の近さという形でグラフ化することができた。特に、名詞や形状詞においては内容面での類似性が強く表れ、『今昔物語集』と『沙石集』という仏教的様相の強い一群と、『古今著聞集』『宇治拾遺物語』『十訓抄』の世俗説話的な一群、そして『江談抄』『中外抄』『富家語』の一群がはっきりわかれる形になった。このような結果は、これまでの説話研究の内容を踏まえればある程度予想のつくものではあるが、これらのそれぞれがどれほどの近さでの類似性として捉えられるかという距離的な見解を得られたことは、一定の成果と言える。また、形容詞や動詞においては文体的な要素が強く結果に表れ、特に『今昔物語集』と『日本霊異記』の類似性が強く表れた。『江談抄』『中外抄』『富家語』が１グループとなる点においては大きく変わらなかったが、『宇治拾遺物語』『古今著聞集』『十訓抄』『沙石集』の類似性については、品詞により若干の違いが出ることが確認された。

細かい部分についてさらなる検証が必要と判断されたため、これらの内容について期間内に発表するところまでは行きつかなかったが、これまでの文学者の見解と大きく変わらない結果が出た点については、この手法の有意性を示すものと考えられ、また、それらが単なる感覚的なものとしてや一部分の検証結果としてではなく、作品全体の使用言語傾向から導き出され、なおかつどの程度の類似性なのかを距離的なもので提示できた点において、意義が見出せるものと考える。

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　0件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 平本留理、蓬莱尚幸、河原井翼	４．巻 43号
２．論文標題 テキストマイニングによる説話文学研究の可能性　―『古今著聞集』巻一、抄入部の検証を中心に―	５．発行年 2018年
３．雑誌名 国語の研究	６．最初と最後の頁 12-21
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 ―

１．著者名 平本留理	４．巻 56
２．論文標題 テキストマイニングによる三大説話集の傾向分析　―x-meansによる分析結果に関する考察―	５．発行年 2021年
３．雑誌名 茨城工業高等専門学校研究彙報	６．最初と最後の頁 31-40
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 ―

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 河原井翼、蓬莱尚幸、平本留理
２．発表標題 中世日本文学に対する自然言語処理
３．学会等名 第17回情報科学技術フォーラム
４．発表年 2018年

１．発表者名 平本留理
２．発表標題 テキストマイニングによる説話集間の関連分析　―三大説話集の解析結果を中心に―
３．学会等名 説話文学会
４．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	蓬莱　尚幸 (Horai Hisayuki) (80633346)	茨城工業高等専門学校・国際創造工学科・教授 (52101)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関