E-019

中世日本文学に対する自然言語処理
Natural Language Processing for Japanese Mediaeval Literature

河原井 翼† 平本 留理† 蓬莱 尚幸†
Tsubasa Kawarai Ruri Hiramoto Hisayuki Horai

1. はじめに

我々は、自然言語処理技術を用いて、中世日本文学、特に説話文学の作品分類等の研究を行っている。本稿では、その一例として、『古今著聞集』(以下、『著聞集』とする)の説話に関する分析の報告を行う。『著聞集』とは、30 篇 726 話から構成される説話集である。説話文学は多数の説話を集録するアンソロジーの形態をとっているので、他文献からの引用(以下、抄入とする)された説話が含まれている。『著聞集』には 697 話の説話が集録されているが、そのうち 78 話が抄入された説話だと考えられている。しかし、その中には、出典が不明であり抄入に疑義がある説話が 4 話(『著聞集』第三〇話から第三三話)存在する。この疑義がある 4 話について、抄入か抄入でないかを判定させたい(以下、抄入分析とする)。そこで、二値分類を行うことに広く利用されている機械学習の一手法であるサポートベクターマシン(以下、SVM とする)を利用することにした。これを用いて、2 つに分類された教師データを学習し、未知データの疑義がある 4 話の分類を行った。

2. 形態素解析を用いた分析

テキストの特徴は、そのテキストに含まれている単語集合で表せる。すなわち、テキストの違いは、単語の頻度のばらつきととらえられる。もし偏在する単語がテキスト同士に共起するならば、それらのテキストの類似度は高い。それに対して、遍在する単語は共起していても類似しているとは限らない。そこで偏在に基づいた分析を行う。

テキストを単語に分解するため、形態素解析を使用した[1]。古典に対して形態素解析を行うために、解析エンジン「MeCab」と解析辞書「中古和文 UniDic」を用いたツール「和文茶まめ」を利用した。これらのソフトウェアを用いて、テキストを形態素(ここでは近似的に単語として扱う)に分解する。その際、分解した単語には、解析辞書に含まれる品詞等の様々な情報が付加される。なお、単語の表記が同じでも、品詞が異なる単語は、別の単語として区別する。

そして、分解した単語をテキストごとに数え上げて、単語の出現頻度(以下、単語頻度とする)を求める。得られた単語頻度は、テキストの特徴を定量化したベクトルとみなすことが出来る。このようにテキストをベクトルとして表現し、ベクトル空間モデルで扱っていく。

3. 対象データ

『著聞集』のテキストは、国文学研究資料館のホームページに公開されている『日本古典文学大系 84 古今著聞集』を使用した。このテキストには、漢文の訓点を表す「／」や空白を表す「△」等の記号がつけられている。今回は、これらの記号を利用して分析を行わないため取り除いた。

また、句読点やかぎ括弧は、校訂者が読みやすさを高めるために挿入されている。これらの記号については、先行研究より、形態素解析前で取り除く場合と形態素解析後で取り除く場合でも、ほぼ同等の結果が得られている[2]。校訂者の意図が含まれず、より原文に近づけるため、これらの記号も形態素解析前に取り除いた。

さらに、『著聞集』第一六六話は、他の説話と異なり、説話の途中から一部抄入がされている。抄入分析をするために、両方の性質を持つこの説話は扱いにくいため用いない。

よって、取り扱うテキスト(以降、テキストと説話を同義で扱う)は、抄入ではない『著聞集』本来の説話 618 話、他文献から抄入の説話 78 話(疑義がある 4 話も含む)、計 696 話である。

4. 分析

本章では、今回行った分析の流れを説明する。

4.1 単語文書行列

2 章の手順で説話の定量化をし、全説話に対してベクトルを作成し、単語文書行列として一つの行列にまとめた。分析における単語文書行列の概略を図 1 に示す。行は説話、列は単語に対応している。例えば、一行目である第二話は、「内侍」と「見える」という単語がそれぞれ 2 回出現する。また「大師」は、一度も出現しないことを示している。

この手順で得られた単語文書行列の大きさは、696(説話)×12172(単語)となった。また、偏在性を際立たせるために、遍在する単語を不要語として分析対象から外す。不要語は、品詞ごとに決定し、「助詞」「助動詞」「名詞-数詞」とした。また、形態素解析の結果で「未知語」と判断された単語があったが、その説話にのみ出現する場合がほとんどだったため、今回は不要語とし取り除いた。

	内侍	見える	…	大師
第二話	2	2	…	0
第三話	0	1	…	0
…	…	…	⋱	…
第七二六話	0	0	…	1

図 1 単語文書行列の概略

†茨城工業高等専門学校
National Institute of Technology, Ibaraki College

Copyright © 2018 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

4.2 正規化と単語の重みづけ

次に、単語文書行列に対して、正規化と単語の重みづけを行った。

正規化は、各話の単語頻度の総数の差の影響を防止するために行われる。今回、正規化にはベクトルの各単語頻度をベクトルのノルムで除算する方法を用いた[3]。ベクトルのノルムの式は次に示す。

$$\mathrm{Norm(n)} = \sqrt{\sum_{k=1}^{W} (t_{k,n})^2}$$

n は説話、$t_{k,n}$ は単語 k の頻度であり、W は全単語数を表している。この式で説話 n 話での各単語頻度を Norm(n)で除算することにより正規化を行った。

また、TF・IDF 法を用いて単語の重みづけを行った[3]。この重みづけは、説話全体で各単語を評価し、少数にのみ現れる偏在性が高い単語に重みを与える。ある説話での各単語頻度(TF)と逆文書頻度(IDF)の積で求める。用いた式を次に示す。

$$\mathrm{TF \cdot IDF} = \mathrm{TF} \cdot \left(\log\left(\frac{N}{df(t)}\right) + 1\right)$$

df(t)は、全説話の中にある単語 t が出現する説話の数を表し、N は全説話数である。log の底には 10 を採用した。

以上の手法を使い、今回行った抄入分析には、正規化のみの場合と正規化と重みづけを行った場合の 2 通りの単語文書行列を作成し、それぞれ SVM による分類を行った。

4.3 SVM による分析結果

SVM では、カーネル関数を用いて、特徴空間へ写像し、線形分離を行う。写像はカーネル関数で異なるため、分析に適した関数を選択することが望ましい。分析には、分析対象の数に対して特徴の数が大きいデータに有効である RBF カーネルを用いることにした。具体的なツールとしては SVM には Python の機械学習ライブラリ「scikit-learn」の「sklearn.svm.SVC」を用いた。また、この関数には、「class_weigh」を設定することにより、教師データの正例と負例にかける重みの値を調整することが出来る[4]。デフォルトは、重みづけを行わない設定になっているが、教師データの正例と負例の数に大きな差がある場合に用いられる。今回、本来の説話と抄入の説話の数では、大きな差があるため、引数に' balanced'を設定し、重みの自動調整を行った。

正例の教師データである本来の説話と負例の教師データである抄入の説話、計 692 話の教師データと疑義がある 4 話の未知データを用いて SVM で予測を行った。分類した結果を表 1 に示す。疑義がある 4 話は、全て他文献から抄入された説話と分類された。

表 1 SVM による疑義がある 4 話の分類結果

分析手法(SVM)	分析結果			
単語文書行列の調整	第三〇話	第三一話	第三二話	第三三話
正規化	抄入	抄入	抄入	抄入
正規化と重みづけ	抄入	抄入	抄入	抄入

5. 分析結果の検証

本手法の妥当性を評価するために、「leave-one-out 交差検証」を行った。この検証では分類がはっきりとしている 692 話を用いるため、疑義がある 4 話は用いない。また、4 章の分析では RBF カーネルを用いていたが、比較対象として線形カーネルも用いることにした。評価結果は、表 2 に示す。分類精度は、次のように算出している。

$$\text{分類精度[\%]} = \frac{\text{成功した数}}{\text{成功した数} + \text{失敗した数}} \cdot 100$$

ここでの成功した数および失敗した数とは、元の分類と同じ結果が SVM により返されたか否かである。

表 2 より説話全体の分類精度の結果が、ほぼ 90%に達しているため、単語文書行列と SVM を用いた分析は、抄入分析に有効である。本分析で用いた RBF カーネルについては、分類精度が 100%であるので、本分析の結果は妥当である。

線形カーネルついては、抄入の説話を正しく分類することが出来なかった。このことから、分析の対象の数に対して特徴の数が大きいときに RBF カーネルが適していると再確認できた。

以上の結果より、本研究の妥当性が示された。

表 2 SVM の精度評価結果

分析手法(SVM)		本来の説話 (618話)	抄入の説話 (74話)	説話全体 (692話)
単語文書行列の調整	カーネル関数	分類精度[%]	分類精度[%]	分類精度[%]
正規化	RBFカーネル	100	100	100
正規化と重みづけ		100	100	100
正規化	線形カーネル	98.9	28.4	91.3
正規化と重みづけ		99.2	17.6	89.7

6. おわりに

本稿では、SVM を用いて疑義がある 4 つの説話の分類を行った。その結果は全て抄入を示し、分類の妥当性を「leave-one-out 交差検証」を用いて示すことができた。今後、抄入分析を進めるために、本稿とは異なった行列の作成や他の分析手法を使うことで、疑義がある 4 話についてさらに分析を行っていく。

そして本研究は、自然言語処理技術の日本文学分野への可能性を示している。今後も、説話集の巻ごとの特徴や作品に描かれる時代的特徴の分析を行っていく。

参考文献

[1] 奥村 学, “自然言語処理の基礎”, コロナ社, (2010).
[2] 平本 留理, 蓬莱 尚幸, 河原井 翼, “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”, 国語の研究, vol.43, pp.12 – 21, (2018).
[3] Ellis David., “情報検索論―認知的アプローチへの展望”, 丸善 ,(1994)
[4] “SVM: Separating hyperplane for unbalanced classes”, scikit-learn, http://scikit-learn.org/stable/auto_examples/svm/plot_separating_hyperplane_unbalanced.html, (参照:2018-6-28).

Copyright © 2018 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.