深層学習による翻訳文学の書き手の同定に関する研究

2017SS001 安部沙桜里
指導教員：松田眞一

1　はじめに

ある言語で書かれた文学作品を，他の言語に移し換えた作品のことを翻訳文学と言う．私は趣味で，同一の海外文学作品をもとにした複数の翻訳文学作品を読むことがあるが，同じ作品でも翻訳者によって物語全体のイメージが変わる場合があった．この差異に興味を持ち，研究の題材にすることにした．本研究は，翻訳された小説作品における書き手の同定という観点で解析を行う．

2　データについて

2.1　用いるデータ

本研究では，ある特定の海外の文豪の作品に対して，日本の複数の翻訳者が，同一の作品を翻訳している必要がある．この条件にかなう作品として，以下の翻訳作品を扱った．

表 1　扱う作品リスト

原作者	原題	翻訳者	翻訳作品名
Hans Christian Andersen	Den lille Havfrue	楠山正雄	人魚のひいさま
		矢崎源九郎	人魚の姫
	Grantræet	楠山正雄	もみの木
		矢崎源九郎	モミの木
Arthur Conan Doyle	A Study in Scarlet	延原謙	緋色の研究 （第一部第一章～第二章， 第二部第八章～第九章， 第二部第十三章～第十四章）
		大久保ゆう	緋のエチュード （第一部第一章～第二章， 第二部第一章～第二章， 第二部第六章～第七章）
		寺本あきら	緋色の研究 （第一部第一章～第二章， 第二部第一章～第二章， 第二部第六章～第七章）

このうち，延原謙が翻訳した『緋色の研究』は，新潮社より出版されている文庫本での文書 [3] を用いた．また，寺本あきらが翻訳した『緋色の研究』は，Web 上で公開しているサイト「コンプリート・シャーロック・ホームズ」[8] の文書を用いた．それ以外の作品は青空文庫 [2] よりダウンロードした．

2.2　データの処理

書誌状態の作品は書誌をイメージスキャナーで取り込み，光学文字認識（OCR）を用いて，テキストデータに変換した．

変換したテキストデータやダウンロードしたテキストファイルには，ルビや注釈など文章の解析には不要な情報が含まれているため，これらを削除する文章のクリーニング作業を行った．本研究では，ルビの削除に Web[1] より入手可能である自動ルビ削除プログラム「delruby.exe」を用いた．タイトルや注釈など，その他不要な情報は手作業で削除した．

また，解析対象のテキストデータは，原文情報に一致するように目視で確認しながら，1000 字を目安にして区切ってから用いた．基準としたのは，翻訳されたのがそれぞれ最も古い翻訳作品とした．アンデルセンの作品が楠山の翻訳版，コナン・ドイルの作品が延原の翻訳版である．

日本語の文章を統計的に解析するには，文章情報を解析可能な数値データへと変換する必要があるため，本研究では，コンピュータによって判別を行う，自然言語処理技術の形態素解析を用いた．これには，形態素解析フリーソフト MeCab を，統計解析ソフト R 上で実行することができる RMeCab を用いた．（石田 [4] 参照）

2.3　変数

金 [6, 7] において，品詞の n-gram 分布，読点前の文字の分布などに書き手の特徴が現れると示されており，渡邊・松田 [10] でも書き手の同定における変数としてこれらを扱っている．本研究では同様にして，品詞の n-gram 分布と読点前の文字の分布を変数として扱う．テキストデータの各変数については，相対頻度を用いた．

- n-gram 分布
  n-gram とは文字あるいは形態素，または品詞が n 個繋がった組み合わせにて表されるものである．本研究では品詞同士の繋がりの情報をデータとして用いるため，$n = 2$ である bi-gram にて表されるものを変数として扱う．
- 読点前の文字の分布
  読点前の文字の分布は，読点「、」の前の文字の出現頻度を総数で割ったものである．

3　分析方法

分析方法には，深層学習を用いる．交差検証には Leave One Out 法を用いた．（渡邊・松田 [10]，北 [5] 参照）

深層学習を行うにあたっては，統計ソフト R の ‘h2o’ パッケージ [9] を用いて実装する．学習回数は 1000 回とし，活性化関数は Rectifier を用い，その他の各種パラメータはデフォルトのままで検証を行った．

4　分析結果

4.1　n-gram 分布

アンデルセンの作品における翻訳者 2 人の合計 41 データ，コナン・ドイルの作品における翻訳者 3 人の合計 150 データに対して n-gram 分布の項目を集計した．これらの項目のうち，すべてのサンプルに対して出現回数を半数以上と基準を設け，それ未満の項目についてはその他の項目にまとめて集計した．これにより選出した n-gram 分布の

項目は，その他の項目を含めて 31 項目となった．

これらの項目による分析結果は，それぞれ表 2，表 3 の通りである．なお，コナン・ドイルの作品は翻訳者が 3 人となるため，これ以降の結果は，正解数と正解率のみを示す．

表 2　アンデルセンの作品
n-gram 分布の正解率

	楠山	矢崎	正解率
楠山	39	2	0.951
矢崎	16	25	0.610
合計	55	27	0.778

表 3　コナン・ドイルの作品
n-gram 分布の正解率

	正解数	正解率
延原	41	0.820
大久保	21	0.420
寺本	34	0.680
合計	96	0.640

4.2　読点前の文字の分布

4.1 と同様に，読点前の文字の分布を集計し，項目の選出をおこなった．アンデルセンの作品はすべてのデータに対して出現回数 20 以上と基準を設け，それ未満の文字についてはその他の項目にまとめた．コナン・ドイルの作品は金 [6] を一部参考に項目を選出し，それ以外の文字についてはその他の項目にまとめた．これにより選出した読点前の文字の分布は，その他の項目を含めて，アンデルセンの作品が 27 項目，コナン・ドイルの作品が 25 項目となった．

これらの項目による分析結果は，それぞれ表 4，表 5 の通りである．

表 4　アンデルセンの作品
読点前の文字の分布の正解率

	楠山	矢崎	正解率
楠山	35	6	0.854
矢崎	14	27	0.659
合計	49	33	0.753

表 5　コナン・ドイルの作品
読点前の文字の分布の正解率

	正解数	正解率
延原	39	0.780
大久保	33	0.660
寺本	22	0.440
合計	94	0.627

4.3　合同解析

4.1 および 4.2 の結果より，翻訳者によってはいずれかの分布を用いた解析の方が，判別精度が高くなることがあった．これより，n-gram 分布と読点前の文字の分布から集計した項目を合わせて解析を行った．4.1 および 4.2 にて選出した項目，それぞれ合計 58 項目，56 項目に対する分析結果は，表 6，表 7 の通りである．

表 6　アンデルセンの作品
合同解析の正解率

	楠山	矢崎	正解率
楠山	39	2	0.951
矢崎	15	26	0.634
合計	54	28	0.802

表 7　コナン・ドイルの作品
合同解析の正解率

	正解数	正解率
延原	38	0.760
大久保	33	0.660
寺本	36	0.720
合計	107	0.713

5　まとめ

アンデルセンの翻訳作品に対する解析は，いずれの場合も 0.75 以上の判別精度を示したが，1 人の翻訳者に比べてもう 1 人の検出には精度が劣っていた．コナン・ドイルの翻訳作品に対する解析は，いずれの場合も 0.6 以上の判別精度を示したが，n-gram 分布の解析において 3 人の翻訳者を個別に見た場合は，0.4 以上の差を示した．

この結果から，翻訳文学は一定の精度で書き手の同定が可能ではあるが，高い分離ができない要因があると考えらえる．特に，本研究の題材が，同一の海外文学作品を基にして翻訳された文章であることから，文章のニュアンスが似てしまい，形態素解析によるデータに類似性が生まれる可能性がある．

また，いずれの作者の作品においても，n-gram 分布と読点前の文字の分布の全体の判別精度を比べると，n-gram 分布の方が高い判別精度を示した．しかし，翻訳者別に見た場合では読点前の文字の分布の方が精度が高いを示す場合もあった．4.3 から示される通り，合同解析では最も高い判別精度を示す結果となった．これより，翻訳者によっては，n-gram 分布または読点前の文字の分布のいずれかに，より特色が現れることが考えられる．

6　おわりに

本研究より，深層学習を用いる結果として高い判別精度ではないが，翻訳文学は一定の精度で書き手の同定が可能であることが分かった．今回，文学を統計学の視点から見ることで新たな発見があった．今後，翻訳文学を含む文学作品を読む際は，今回の研究を踏まえてみたいと思う．

参考文献

[1] AOKIDS Home Page：青空文庫のテキストからルビを削除するには，http://www.aokids.jp/others/delruby.html (2020/7 閲覧)

[2] 青空文庫：http://www.aozora.gr.jp/ (2020/7 閲覧)

[3] A．コナン・ドイル (延原謙 訳)：『緋色の研究』，新潮社，1953.

[4] 石田基広：『R によるテキストマイニング入門（第 2 版)』，森北出版，2017.

[5] 北栄輔：『R で学ぶデータサイエンス　―データマイニングの基礎から深層学習まで―』，オーム社，2018.

[6] 金明哲：読点の情報に基づく文献の分類，情報処理学会『全国大会講演論文集』第 46 回人工知能及び認知科学，131-132，1993.

[7] 金明哲：分節パターンに基づいた文書の書き手の識別，『行動計量学』40(1)，17-28，2013.

[8] 寺本あきら：コンプリート・シャーロック・ホームズ，https://221b.jp/ (2020/9 閲覧)

[9] Package‘h2o’：https://cran.r-project.org/web/packages/h2o/h2o.pdf (2020/6 閲覧)

[10] 渡邊翔・松田眞一：『深層学習を用いた文章の書き手の同定』，南山大学紀要『アカデミア』理工学編，18，1-13，2018.