# 英米文学の原書書評に基づく日本語訳本の売上予測

DENG JUNFU　　掛谷英紀
筑波大学システム情報工学研究科
s1920870@u.tsukuba.ac.jp　　kake@iit.tsukuba.ac.jp

## 1 はじめに

日本の出版業界の市場規模が 1997 年から年々縮小し、「出版不況」といわれる長期低迷の状況に陥っている[1]。その影響を受け、書店と印刷業全体が右肩下がりになり[2]、印税収入に依存する出版翻訳業者も厳しい境地に晒されている[3]。

出版翻訳には、海外で出版された書籍の情報を収集する「原書探し」や、作品の内容を要約し編集会議で吟味する作業などが必要であった[4]。海外書評サイトでの書籍概要とユーザーレビューなどの既存情報を活用し、日本で売れるパターンとそうでないパターンを把握できれば、刊行までの時間が大幅に短縮し、出版のリスクも低減すると予想される。

近年、書評を対象とした分類研究が盛んに行われている。例えば、橋本ら[5]が最大エントロピー法を用いたイデオロギー分析や、掛谷ら[6]がアマゾンのユーザーレビューに基づいた先見性の分析、樋口ら[7]が書評を用いた政治的見解の異なる人物像の特徴分析などがある。

本研究では既存の分類モデルを参考し、和訳された英米文学作品の売れ行きを予測するシステムの実現を目的とする。具体的には、紀伊国屋書店 [8]で掲載された英米文学作品を対象に、海外最大級の書評サイト GoodReads[9]の情報から対象作品が日本で売れるかどうかを分類するモデルを構築する。その上、売れる・売れない作品の特徴を分析する。

## 2 提案システム

提案システムの全体構成を図 1 に示す。まず、紀伊国屋の英米文学ジャンルから和訳本のリストを抽出し、研究対象の選定基準に準ずるかどうかを判断する。対象外の作品は除外し、対象となる作品については GoodReads から原書の情報を抽出する。

原書の情報は発売日、レビュー数など基本情報（構造的データ）と概要、レビューのテキストデータがあるが、概要のみある作品とレビューのみある作品があるため、実験対象を概要群とレビュー群、2 つの実験群に分けて、複数のモデルを用いて売れ行きの分類実験を行う。

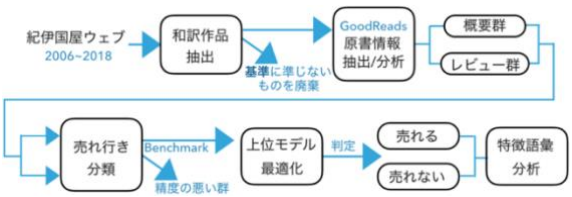


**図 1　提案システムの構成**

実験結果より、平均精度のよい実験群の上位モデルを選出し、最適ハイパーパラメーターを探して再実験する。最後は売れ行きのよい作品とそうでない作品の特徴的語彙を抽出し分析する流れである。

### 2.1 分析対象の選定基準とラベリング

分析対象として、1)海外で出版され、その後和訳された文学作品、2)名作の再版、新訳本ではないこと、3)日本で発売されてから一年以上経った作品の三つの基準に全て合致するものを選んだ。名作については、Wikipedia の「もっとも多く翻訳された著作物」[10]と Norwegian の「Top 100 Books of All Time」[11]、2 つのリストに記載される作品と定義する。上述した基準に基づき、紀伊国屋書店から 2018 年 12 月前に出版された 900 作品を抽出し、GoodReads から基本情報と概要のある原書 756 点、レビューのある原書 621 点（和訳本が出版される前のもの）を収集した。

売れ行きの良しあしをラベリングする際に、本の実売部数を基準とするのが一番理想的だが、そのデータは出版社しか分からず、一般人として入手するのが極めて困難だと考えられる[12]。そのため本研究は日本最大級の書評サイト BookMater[13]から収集した「登録者数」を「注目度」と定義する（登録者数は発売年数の増加に比例する可能性がため両者間の相関関係を調べたが、0.17 しかない）。それを基準として、注目度分布の上位 3 割を売れる・下位 3 割を売れない、のようにラベリングした。

## 2.2 書籍関連情報の相関分析

原書の基本情報となるレビュー数、出版されたバージョン数、平均評価（星）、ページ数とレビュー数も売れ行きと関連する可能性があるため、各実験群の基本情報と登録者数との相関関係を分析した。その結果、売れ行きの基準となる登録者数との相関係数が最も高いのは原書のレビュー数であるが、概要群 0.30、レビュー群 0.33 しかないため、基本情報のみで和訳後の売れ行きを予測するのは難しいと考えられる。

## 2.3 機械学習の手法

2.1 の基準でラベリングしたサンプルは、概要群 452 点（売れる 225、売れない 227、平均単語数 72、標準偏差 137）、レビュー群 332 点（売れる 181、売れない 151、平均単語数 3894、標準偏差 6323）である（レビュー者数が 5 以下の作品は除外）。実験は多種類の学習器と組み合わせる形で行った（図 2）。

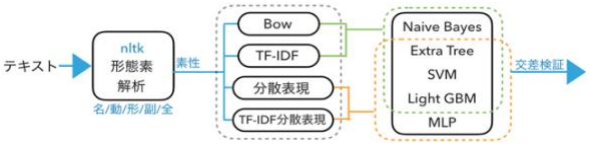


**図 2 実験手法の詳細**

具体的には、まずテキストから記号と符号を除去し、spacy[14]で形態素解析を行う。各種素性の深層格を抽出し、全素性、名詞、動詞、形容詞と副詞の 5 つのパターンに分けて実験する。選択した素性を BoW(Bag of Words)と GloVe[15]で作成された分散表現でベクトル化する。単語の重要度および稀少性の影響も調べたいため、埋め込む際に tf-idf 法（または idf 法のみ）で処理したパターンとそうでないパターンに分けて行う。分散表現の場合は、テキストにある全素性の分散表現の平均を取り、1 作品あたり 300 次元の特徴量ベクトルを作成する。文書 $d$ の分散表現ベクトル$\boldsymbol{v}(d)$は

$$\boldsymbol{v}(d) = \frac{1}{n}\sum_i^n \boldsymbol{w}(w_i) \qquad (1)$$

と表される。tf-idf を重みとしてつける場合は、原田ら[16]の手法に基づき

$$\boldsymbol{u}(d) = \frac{1}{n}\sum_i^n tfidf(w_i)\cdot \boldsymbol{w}(w_i) \qquad (2)$$

を用いる。

実験は、ナイーブベイズ、サポートベクターマシン(SVM)、ExtraTrees（ランダム木）、LightGBM、多層パーセプトロン（MLP）の 5 種類の学習器と指定素性のテキストベクトルを図 2 のように組み合わせて行う。できるだけ早く相性の良い組み合わせを選出するため、すべての学習器はデフォルトのパラメータで学習した。

# 3 実験

## 3.1 クロスバリデーションの結果

各学習法のクロスバリデーションの結果を表 1 と表 2 に示す。ここでリストアップしたのは、それぞれの素性で実験する際に、精度の一番良いモデルである。概要群について、全素性を使い、BoW で化して Naive Bayes で行った実験は最高 70.7%の精度を得た。他の素性のみ使った場合も 60%以上の結果が得られた。

**表 1 概要群の実験結果(%)**

| ペア | 素性 | モデル | 再現率 | 適合率 | F 値 |
|---|---|---|---|---|---|
| 1 | 名詞 | Naive Bayes(BoW) | 67.4 | 70.0 | 66.4 |
| 2 | 動詞 | ExtraTree(TFIDF) | 64.7 | 65.3 | 65.0 |
| 3 | 形容詞 | Naive Bayes(BoW) | 64.2 | 64.2 | 64.2 |
| 4 | 副詞 | ExtraTree(TFIDF) | 60.8 | 60.8 | 60.8 |
| 5 | 全素性 | Naive Bayes(BoW) | 70.7 | 70.8 | 70.7 |

**表 2 レビュー群の実験結果(%)**

| ペア | 素性 | モデル | 再現率 | 適合率 | F 値 |
|---|---|---|---|---|---|
| 6 | 名詞 | SVM(TFIDF) | 82.5 | 82.4 | 82.4 |
| 7 | 動詞 | ExtraTree(GloVe-IDF) | 72.4 | 74.2 | 72.5 |
| 8 | 形容詞 | LGBM(GloVe-IDF) | 72.5 | 72.7 | 72.5 |
| 9 | 副詞 | SVM(TFIDF) | 71.3 | 72.6 | 71.4 |
| 10 | 全素性 | ExtraTree(GloVe-IDF) | 79.2 | 80.0 | 79.2 |

レビュー群については tfidf と SVM の組み合わせが一番良く、名詞のみ使った場合 82.5%の精度に達した。他の素性のみ使った場合の精度も概要群より高く、70%以上の結果が得られた。結果から原書の海外レビューから和訳本の売れ行きを予測する可能性があると考えられ、売れ行きを左右する特徴は名詞、形容詞、副詞などに潜んでいることがわかる。

## 3.2 上位モデルのパラメータ最適化

概要群とレビュー群の分類実験で精度の良かった

素性と分類器の組み合わせに対して、パラメータ最適化を行った。具体的な手法を図 3 に示す。

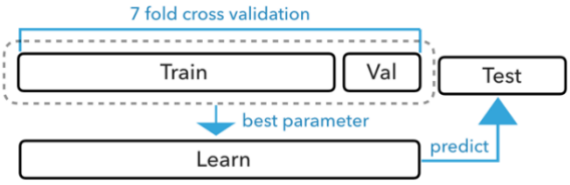


**図 3 パラメータ最適化の手法**

まずすべてのデータを学習用 75%、テスト用 25%の割合で分割し、学習用データのみ使って探索する。次に 7 分割のクロスバリデーションで平均精度の一番良いパラメータの組み合わせを選出する。最後は選出したパラメータで設定し、探索用のすべてのデータを学習してテストデータで検証する。上記の手法で実験した結果、ペア 10 のモデルのパラメータを(n_estimators = 245, max_depth = 16, max_feature = 17)で設定した場合、85.4%の f1 値が得られました。表 3 と図 4 は実験の混同行列と信頼度曲線である。

**表 3 混同行列**

| | 売れる | 売れない | 再現率 | 適合率 |
|---|---|---|---|---|
| 売れる | 39 | 6 | 86.6 | 86.6 |
| 売れない | 6 | 36 | 85.7 | 85.7 |

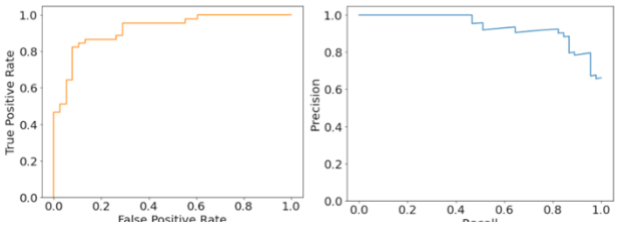


**図 4 ROC 曲線と PR 曲線（AUC 値 0.94/0.92）**

売れる・売れない作品のいずれに対しても、85%以上の適合率と再現率が得られた。また、ROC 曲線から見ても、偽陽性率が低い時でも高い真陽性率が見られるため、一般的に言うと、信頼性の高いモデルと考えられる。

## 3.3 登録者数を予測する実験

以上の実験でにおいて、翻訳小説の売れ行きを最高 85.4%で判断することができるが、どれほど売れる・売れないのかという問題がある。より有益な予測を行うために、多層ニューラルネットワークを用いて、GoodReads のレビューに基づいて BookMeter の登録者数を直接予測することを試みた。実験に用いたネットワークは、入力/出力層と 3 層の隠れユニットで構成されている。各隠れユニットは全結合層、バッチ正則化層、dropout 層で接続され、最終的に ReLU 関数で活性化する。各層のニューロン数と係数を図 5 に示す。

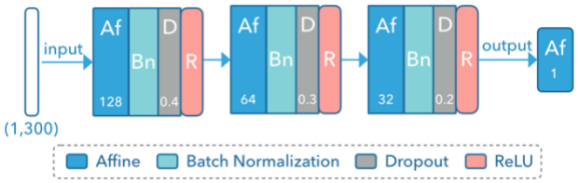


**図 5 ニューラルネットワークの構成**

Optimizer は RMSprop で、損失関数は平均二乗誤差（MSE）、評価関数は平均絶対誤差（MAE）を用いた。全 epoch 数は 500、バッチサイズ 40 で学習し、学習率$lr(t)$は式 3 のように、100 エポック以後徐々に減衰させる。なお、early stopping は 40 エポックに設定する。

実験は 3.2 で高い精度に達する idf 付き Glove 分散表現のベクトルを使う。登録者数の分布が偏っているため、対数変換の手法で前処理する。具体的な手順は図 6 のように、まず 562 個のデータを訓練データとテストデータに分割し、その比率を 7:3 とする。次に訓練データを 7 分割のクロスバリデーションで学習する。その都度、訓練されたモデルを用いてテストデータを予測する。最後は、7 つの予測値の平均値を計算し、最終の予測結果としてモデルを評価する。

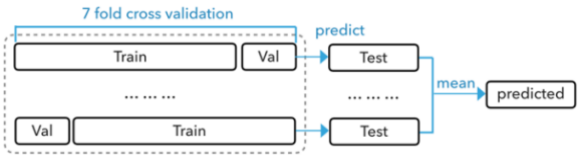


**図 6 予測値を出す手順**

提案モデルは、GoodReads のレビュー数を用いて、線が原点を通る条件のもとで作成した線形回帰モデルと比較し、ニューラルネットワークと線形回帰モデルの予測値から与えられる RMSE、MAE、相関係数を用いて評価する。結果は表 4 に示す。

提案モデルの予測値と実際の登録者数との相関は 0.50 であり、線形回帰で与えられた値より 31%高い結果が得られた。提案モデルと線形回帰による真値と予測値の比較を図 7 に示す（登録者数が対数の形でプロットされている）。

表 4　予測結果の評価

| データ | RMSE | MAE | 相関係数 |
|---|---|---|---|
| Neural Network | 1.79 | 1.21 | 0.50 |
| Linear Regression | 3.55 | 3.06 | 0.19 |

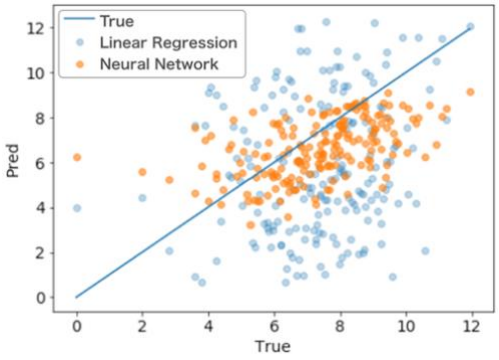


図 7 予測値の真値

図に示されているように、ニューラルネットワークの方が線形回帰より良い予測値を与えていることが分かる。

## 5 考察

売れる・売れないとラベリングされた作品にどのような特徴があるかを調べるため、レビューテキストの単語を出現確率

$$\rho_j^{(i)} = \frac{c_j^{(i)}}{\Sigma_{k=0}^{1} c_k^{(i)}} \tag{3}$$

の順にソートした。ここでの$c_j^{(i)}$は、単語$i$がクラス$i$での出現回数で、$c_j^{(i)} \geq n$ を満たすように計算する。$n$は名詞の語彙数を 100 とした時、各素性の語彙数に比例する。つまり出現頻度に偏りの大きく、しかも一定回数以上出現した単語のみリストアップした。これらを表 5 に示す。

結果から刺激性の強いサスペンス(killer, detective)系、スリラー(thriller)系及び奇想天外(fantasy, quirky)のものと、SF(dystopia, universe, postapocalyptic)作品が特に愛読されていることがわかる。

一方、英米文化の色濃い作品(church, slavery)や、詩(poems)、生活感(allergies, students, coffee)の強い作品は人気がないように見える。国や地域、軍隊など言葉が出たが、元のテキストから確認したところ、欧米の戦争や歴史、移民などを語る物語ということがわかるが、それらも日本文化に距離があるため売れ行きが芳ばしくないと考えられる。

表 5　レビュー群の実験結果(%)

| 売れる作品 | | 売れない作品 | |
|---|---|---|---|
| circus | 96% | allergies | 100% |
| genetic | 91% | Haitian | 98% |
| dystopian | 90% | ballet | 99% |
| surreal | 90% | cricket | 99% |
| postapocalyptic | 85% | expedition | 97% |
| killer | 85% | coffee | 87% |
| travel | 85% | poems | 87% |
| hotel | 82% | church | 86% |
| thriller | 81% | animals | 84% |
| universe | 80% | poem | 84% |
| realism | 79% | translation | 83% |
| quirky | 78% | memoir | 73% |
| detective | 77% | slavery | 72% |
| imagination | 75% | students | 72% |
| fantasy | 74% | account | 72% |

## 6 おわりに

本研究は、海外書評サイトの英語レビューに基づき、和訳本の売れ行きを予測するシステムを考案し、売れる・売れないものの二値分類実験を行った。実験は概要群とレビュー群に分け、それぞれ全素性、名詞、動詞、形容詞と副詞で分類した。結果として概要群では最高 70%、レビュー群では最高 82%の精度が得られた。パラメータ調整後にレビュー群では最高 85.4%の精度が得られ、信頼度曲線から見てこのモデルは信頼できると考えられる。

次に、レビュー群のみを対象とし、登録者数の値を予測するモデルを提案した。ニューラルネットワークを用いた場合、予測値と真値の相関係数は 50%で、線形回帰より 31 ポイント高い結果が得られた。

さらに、売れる作品と売れない作品のレビューに頻出する単語を調査し、刺激性の強い作品、非現実・SF の要素のある作品が特に日本で売れており、英米文化の色濃い作品やポエムは人気ではないことが明らかになった。実験対象が日本と海外での売れ行きの差異を調べ、両方でも売れる作品とそうでない作品、そして日本・海外のみ売れる作品はそれぞれ違う特徴を持っていることが明らかになった。

## 参考文献


1. 出版科学研究所,“日本の出版販売額”, 出版指標年報 2020 年版, 2020.
2. 日本印刷産業連合会, “印刷産業の出荷高・事業所数・従業員数の推移”, マーケティング・データ・ブック 2020, Vol.18, 2020.
3. 山岡洋一, “翻訳の過去・現在・未来”, 「翻訳通信」100 号記念セミナー, 2010.
4. 株式会社トランネット, “出版翻訳とは”, https://www.trannet.co.jp/about-publish-translation.html, 2020.
5. 橋本ら, “Web 上のレビュー記事のイデオロギー分析とその応用”, 言語処理学会第 16 回年次大会, pp.740-743, 2010.
6. 掛谷ら, “書籍のレビューに基づく先見性のある人物の特徴分析”, 言語処理学会第 22 回年次大会, pp.1149-1152, 2016.
7. 樋口ら, “国会会議録と書評を用いた政治的見解が異なる人物像の分析”, 2019 年度日本選挙学会, 2019.
8. 紀伊国屋書店, “英米文学ジャンル”, www.kinokuniya.co.jp/f/dsd-101001015025001--
9. GoodReads , www.goodreads.com.
10. Wikipedia, “もっとも多く翻訳された著作物”, 2020.
11. Norwegian Book Clubs, “The top 100 books of all time”, 2002.
12. 上原龍一, “本の実売部数の調べ方”, www.uehararyuichi.com/jitsubai-299
13. BookMeter, www.bookmeter.com
14. spacy, www.spacy.io/
15. Jeffrey Pennington et al, “GloVe: Global Vectors for Word Representation” , Proceedings of EMNLP 2014, pp.1532-1543, 2014.
16. 原田ら, “単語の分散表現および if-idf 法を用いた自動ようやくシステム”, ARG WI2, No.9, pp49-50, 2016.