

子どもの選好を考慮した類似絵本検索システムの基礎検討

—絵本に関するレビュー文の構成分類—

宮本 華奈^{†,††} 松下 光範[†] 服部 正嗣^{††}

[†] 関西大学総合情報学部 〒569-1095 大阪府高槻市霊仙寺町 2-1-1

^{††} 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所 〒619-0237 京都府相楽郡精華町光台 2-4

E-mail: [†]{k794670,t080164}@kansai-u.ac.jp, ^{††}hattori.takashi@lab.ntt.co.jp

あらまし 本研究の目的は、子どもの選好を考慮した類似絵本検索システムの構築である。養育者が子どもに読み聞かせる絵本を選択する際、オンラインショッピングサイトや国内の絵本・児童書情報サイトの検索機能が用いられているが、この方法では子どもの興味や選好を考慮した絵本を検索することが難しい。そこで、本研究では絵本のレビューを用い、絵本に対する子どもの選好の収集を試みる。一般的に絵本のレビューは読み聞かせをする養育者によって投稿されており、子どもが実際に取った行動や発言した内容についても記述されている。そのため、絵本のレビューは子どもの選好に関する情報の取得に適していると考えられる。しかし、これらのレビューには養育者の主観も混在しているため、子どもの選好を効率よく得るには養育者の主観と子どもの反応を区別して扱う必要がある。本稿では、機械学習を用いてレビュー中に含まれる要素を自動分類する手法について検討する。

キーワード レビュー分類, 絵本検索, Random Forest, Support Vector Machine

Toward review search system considering children's taste

—Classifying review texts for picture books—

Haruna MIYAMOTO^{†,††}, Mitsunori MATSUSHITA[†], and Takashi HATTORI^{††}

[†] Faculty of Informatics, Kansai University 2-1-1 Ryozenjicho, Takatsuki, Osaka 569-1095 Japan

^{††} NTT CS labs., NTT Corp. 2-4 Hikaridai, Seikacho, Sorakugun, Kyoto 619-0237 Japan

E-mail: [†]{k794670,t080164}@kansai-u.ac.jp, ^{††}hattori.takashi@lab.ntt.co.jp

Abstract The purpose of this study is to construct a similar picture book retrieval system considering children's taste. When caregiver selects picture book to read to children, they often use Web services such as search function of online shopping site, domestic picture book site, and children's book site. But these sites do not provide search results by taking children's interests into account. To solve the problem, this paper tries to collect children's taste for picture books from review texts. Since a picture book review is written by a caregiver who makes reading story, in general, it includes the action actually taken by children and the contents which they said. This means that a picture book review will be a suitable medium to acquire information on children's taste. To extract children's taste from the reviews effectively, it is necessary to distinguish the subjective comments by the caregiver and the reaction of the child. To meet this goal, this paper investigates whether automatic classification of sentences in picture book reviews is possible or not, by using manually tagged review sentences.

Key words review classification, picture book search, Random Forest, Support Vector Machine

1. はじめに

現在、絵本の新刊は毎年1,500冊以上出版されている[1]。これら大量の絵本から養育者が子どもに読み聞かせる絵本を選択

する際、オンラインショッピングサイトや国内の絵本・児童書情報サイトの検索機能が用いられているが、現在提供されている機能では興味や選好に基づいて絵本を検索することが難しい。特に、大人が好む本と子どもの好む本には大きなずれがあるこ

とが指摘されており [2], 現状では子どもの興味や選好を考慮した絵本を検索することはできない。

こうした背景の下, 本研究では絵本に対する子どもの選好に着目した絵本検索システムの構築を目指す。この足がかりとして, 本稿では絵本のレビューを用い, 絵本に対する子どもの選好の収集を試みる。一般的に絵本のレビューは読み聞かせをする養育者によって投稿されており, 子どもが実際に取った行動や発言した内容についても記述されている。そのため, 絵本のレビューは子どもの選好に関する情報の取得に適していると考えられる。しかし, これらのレビューには養育者の主観も混在しているため, 子どもの選好を効率よく得るには養育者の主観と子どもの反応を区別して扱う必要がある。そこで本稿では, 機械学習を用いてレビュー中に含まれる要素を自動分類する手法について検討する。

2. 関連研究

2.1 子どもの選好の取得方法に関する研究

松村ら [3] は, 読み聞かせ時の子どもの反応の記録から, 子どもの選好を取得する方法について検討している。読み聞かせ時の子どもの反応を「言語」, 「集中」, 「遊び」, 「主張」, 「感情」, 「読解」の6つに分類し, それぞれを平均値と比較することで子どもごとの選好の特徴を取得している。しかし, この方法は毎回読み聞かせをした時に子どもの反応を記録しなければならず, 読み聞かせ中の子どもの反応を覚えておく必要があるため, 養育者にとって負担となる。

湯川ら [4] は, WEB カメラとタッチパネルディスプレイを用い, 電子書籍を用いて絵本の読み聞かせを行う場合の子どもの興味度を自動検出できることを示している。この研究では, 絵本の読み聞かせ中の様子を WEB カメラで取得し, 動画画像解析を行うことにより得られた視線方向や顔移動量から絵本に対する子どもの興味度を推定している。しかし, この方法は WEB カメラが使用できるタブレット端末等の絵本の読み聞かせに限られており, 紙の絵本では興味度の推定を行うことができない。

今満ら [5] は, 幼稚園のお話会のような, 集団に対する読み聞かせのための絵本選択支援について研究している。この研究では, お話会の記録より子どもの反応を収集し, 子どもの反応付きの絵本リストを作成している。また, この絵本リストを用いて実際に絵本選択を行った結果, 子どもの反応を用いることが絵本選択支援に有効であることを明らかにしている。本研究では, 子どもの反応をレビューから取得することにより, 多様な子どもの反応を取得する。

藤田ら [6] は, 絵本テキストの言葉の難しさや読みやすさなどから子どもの年齢に適した絵本を探すことや, 表紙の画像情報から絵柄の似た絵本を探すことができる絵本検索システムの研究を進めている。しかし, このシステムでは作者やタイトル, 出版年などの書誌情報や絵本の内容情報といったコンテンツベースの検索に限られており, レビューに含まれているような実際の子どもの反応や絵本に対する養育者の評価から絵本を探すことはできない。

松村ら [7] は, 子どもの質問に子どもの興味が現れることに

着目した絵本推薦システムを構築している。この研究では, 養育者が子どもの質問を Web 上に登録し, その質問を見た他の養育者がおすすめ絵本を紹介することによって絵本推薦を実現している。しかし, この方法では質問とその回答にあたる絵本が揃うまでは時間が必要である。また絵本を推薦しにくい質問に対しては回答数が少数となることから, 子どもの選好を考慮した絵本を選ぶために十分な情報を得ることは難しい。

2.2 レビューからの情報抽出に関する研究

Uehara ら [8] は, 発達心理学の視点より絵本の子どもの認知発達効果が認められていることから, 実際に絵本を読み聞かせた養育者や保護者が絵本ナビに書き込んだ大量のレビューより, 絵本が子どもに発達的な効果を及ぼす様子を明らかにしている。この研究では, 絵本 16 タイトル, 345 レビューを対象として, レビュー中の記述をレビュアーの反応や子どもの反応, 絵本の描写などに人手で類型化している。しかし, 全てのレビューに対して人手で類型化することはデータ作成者にとって負担となる。

山澤ら [9] は, 書き手の立場や観点が様々なレビューの中から, 信頼できる文(有用文)の自動抽出方法を提案している。この研究では, Amazon レビューに対して人手で有用文に分類を付与したのち, 全ての品詞を対象としている。また, 有用文であると判断された文とそうでない文のそれぞれで, 一方での出現頻度が高く, かつもう一方での出現頻度が高い品詞を抽出して自動抽出の精度を上げる方法を提案している。本稿でも単語の出現頻度を見ることで分類の精度向上を試みる。

岩井ら [10] は, ネタバレを防止するため, レビューに含まれるあらすじをそれ以外と分類する手法を提案している。この研究では, Amazon レビューよりストーリーを持つカテゴリのレビューを取得し, あらすじとみられる文に人手で分類を行い, 機械学習で代表的な5つのモデルを用いてあらすじ分類に最適なモデルを明らかにしている。本稿でも従来より自然言語処理に多く使われている [11] Support Vector Machine (以下 SVM と記述) [16] と Random Forests (以下 RF と記述) [17] を用いて絵本のレビュー分類に最適なモデルを明らかにする。

山西ら [14] は, レビュー中にみられる形態的特徴から評価視点の抽出パターンを作成し, 評価視点の抽出を行っている。この研究では, 評価対象についての意見, 感想が記述された文を評価文とし, その中でもレビュアーの意見や感想が述べられている部分を評価視点, 評価対象に対する意見や感想の評価を表す語を評価表現としている。そしてその二つが共に含まれた部分を評価部分と定義し, 抽出対象としている。

高野ら [15] は, レビュー文に含まれる因果関係に着目し, 有用な情報を自動抽出する手法を提案している。この研究では, レビュー文を構文解析することにより, レビューがレビュー対象に関して抱いている評価の自動抽出を可能にしている。

2.3 本研究の位置付け

子どもの選好を取得する方法として, 先行研究では子どもの反応を記録することで子どもの興味や選好を推定する手法が多く挙げられていた。しかし, 毎日の読み聞かせにおいて毎回子どもの反応を記録することは養育者にとって負担となる。また,

表 1 使用したレビューの作品

作品名	著者	出版社
きんぎょがにげた	作: 五味 太郎	福音館書店
はらぺこあおむし	作・絵: エリック・カール 訳: もり ひさし	偕成社
しろくまちゃんのほっとけーき	作: わかやま けん	こぐま社
もこもこもこ	作: 谷川 俊太郎 絵: 元永 定正	文研出版
いないいないばあ	文: 松谷 みよ子 絵: 瀬川 康男	童心社
がたん ごたん がたん ごたん	作: 安西 水丸	福音館書店

表 2 分類のタイプとその定義

分類	定義
(1) 物語に対する感想	物語の内容についての感想や印象に関する文
(2) 絵柄に対する言及	絵柄の雰囲気やタッチについての感想や印象に関する文
(3) 書誌に対する言及	出版社や著者, 出版年, シリーズ, 対象年齢についての感想や印象に関する文
(4) 子どもの反応	子どもの発言や行動に関する文
(5) 物語内容	物語の内容やあらすじ, 登場人物など, 物語を説明する文
(6) 書誌情報	出版社や著者, 出版年, シリーズなど絵本の客観情報に関する文
(7) 入手経路/入手目的	その絵本の入手経路や入手目的が含まれる文
その他	上記の分類に当てはまらないと判断した文

表 3 3人の評価者による分類の例

レビュー文	評価者 1	評価者 2	評価者 3	結果
言葉を覚えたての時期から読んであげると、たくさん子どもとコミュニケーションがとれて養育者子どもも楽しめる絵本だと思います。	4	その他	1	不
いろいろな絵本がありますが、一緒に遊べる絵本なので養育者子どもで楽しめます。	1	4	その他	不
色彩も綺麗だし、いろんなモノが書いてあるページはモノの名前を覚えるのにも役立ちました。	2	3	2	2
カラフルな絵が気に入って購入しました。	7	2	7	7
絵がきれいで、工夫されているので、読み聞かせしていても飽きません。	2	2	2	2
小さいころも大きくなったいまもときどき出してきては子供が読んでいます。	4	4	4	4

先行研究で挙げられている絵本推薦システムについては、絵本選択に有用なシステムと考えられる。しかしこれらのシステムはコンテンツベースの推薦システムのため、子どもの反応に基づいた検索をすることはできない。

本研究では絵本のレビューから様々な子どもの選好を取得することにより、養育者の負担を軽減し、自身の子どもの反応と比較して絵本選択を行うことができるシステムを目指す。また、絵本の内容よりも実際に読み聞かせを行った時の子どもの反応に着目し、絵本選択に有用なレビュー文を取得する。

本研究で扱う絵本のレビューには子どもが実際に取った行動や発言した内容が含まれており、子どもの選好に関する情報の取得に適していると考えられる。しかし、これらのレビューには養育者の主観や内容情報も混在しているため、子どもの選好を効率よく取得することが困難である。

そこで本稿では、絵本のレビュー文に含まれる子どもの反応や絵本のレビューを構成するその他の要素を、機械学習を用いて自動で分類することで、子どもの選好を考慮した絵本探索システムの構築を目指す。機械学習時に学習データおよびテストデータとして用いるため、本研究では先行研究に習い、人手でレビューに対して分類を付与する。機械学習のモデルには先行研究でよく用いられている SVM と RF を用い、どちらが絵本のレビュー文の分類に適しているか検証する。

表 4 分類の割合

分類	分類された文の数	割合
1	672	11.9%
2	307	5.4%
3	158	2.8%
4	2295	40.7%
5	214	3.8%
6	11	0.2%
7	254	4.5%
その他	1728	30.6%

3. 実験

本章では、詳細な分類を可能にするため、3人の評価者によって分類されたデータを用い評価を行った。また、データ中の低頻出及び高頻出な単語を除き、分類精度の向上を図った。

3.1 実験の流れ

絵本情報サイト「絵本ナビ」(<http://www.ehonnaivi.net/>)からレビュー数が上位6作品のレビュー文を取得した。絵本ナビではレビューを投稿する際、絵本に対し5段階の評価を付与することが必須となっており、取得したレビュー文は評価の良い順番に並んでいる。そのため、作品に付与されたレビュー文の中から、句点区切りのレビュー文を1文として、ランダムで

表 5 SVM による分類結果

分類	適合率	再現率	F 値	テストデータ数
1	0.67	0.22	0.33	134
2	0.69	0.51	0.58	61
3	0.65	0.17	0.27	32
4	0.69	0.85	0.76	459
5	0.64	0.05	0.09	43
7	0.74	0.35	0.45	51
その他	0.60	0.76	0.67	348
平均	0.67	0.42	0.45	1128

表 6 RF による分類結果

分類	適合率	再現率	F 値	テストデータ数
1	0.68	0.16	0.26	134
2	0.72	0.50	0.59	61
3	0.48	0.13	0.21	32
4	0.74	0.79	0.76	459
5	0.75	0.09	0.16	43
7	0.84	0.30	0.43	51
その他	0.53	0.82	0.64	348
平均	0.68	0.40	0.44	1128

1 作品あたり 1,000 文 (計 6,000 文) のレビュー文を収集した。次に、絵本レビューに含まれると考えられる要素を (1) 物語に対する感想, (2) 絵柄に対する言及, (3) 書誌に対する言及, (4) 子どもの反応, (5) 物語内容, (6) 書誌情報, (7) 入手経路/入手目的の 7 タイプ及びその他の 8 種類に分類した (表 2 参照)。この分類に基づき、収集したレビュー文がどのタイプに該当するかについて 3 人の評価者により分類を行った。3 人の評価者のうち、2 人以上が同一であった分類をそのレビュー文の分類とした (表 3 参照)。その結果、5,639 文のレビュー文が得られた (表 4 参照)。なお、書誌情報タイプに分類されたレビュー文は他の分類に比べて極端に少なかったため、その他に統合することとした。

分類器に用いる素性は、レビュー文中に出現する名詞、動詞、形容詞 (いずれも原形) とした。上記で得られたレビュー文を Bag-of-Words 法により単語ベクトルとして表し、それらを 2 ラベルの分類に優れた線形分類器である SVM と、複数ラベルの分類に優れた決定木分類器である RF を用いて上記の 8 種類に分類し、分類性能を比較した。

また、SVM ではコストパラメータ、RF では決定木の数などのハイパーパラメータを設定する事でよりデータに合った分類器を作成することができる。各種パラメータを最適化するため、複数のパラメータ候補による事前検討を行った。今回は SVM のコストパラメータとガンマ値の候補に 0.01, 0.1, 1.0 を、カーネルの候補に linear と rbf を用いた。RF では決定木の数の候補に 10, 30, 50, 70, 90, 110, 130, 150 を、各決定木に用いる説明変数の数の候補に auto, sqrt, log2, None を用いた。RF では、決定木の数を調整することで、過学習を防ぐ事ができる。さらに、各決定木に用いる説明変数の数を調整することで、計算コストを削減することができる。

各種パラメータの事前検討の結果、SVM ではコストパラメー

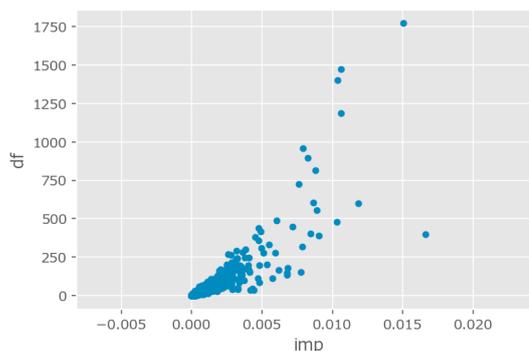


図 1 DF 値と重要度の関係

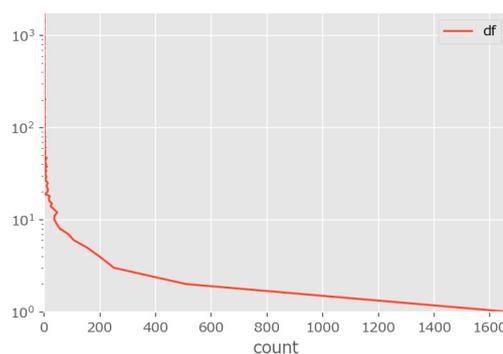


図 2 DF 値の頻度分布

表 7 低頻度語、高頻度語除去後の SVM による分類結果

分類	適合率	再現率	F 値	テストデータ数
1	0.56	0.23	0.33	134
2	0.66	0.57	0.61	61
3	0.76	0.26	0.37	32
4	0.71	0.82	0.76	459
5	0.50	0.03	0.05	43
7	0.73	0.46	0.55	51
その他	0.59	0.77	0.67	348
平均	0.65	0.45	0.48	1128

表 8 低頻度語、高頻度語除去後の RF による分類結果

分類	適合率	再現率	F 値	テストデータ数
1	0.57	0.24	0.34	134
2	0.59	0.66	0.62	61
3	0.67	0.25	0.35	32
4	0.74	0.71	0.72	459
5	0.50	0.10	0.15	43
7	0.66	0.43	0.51	51
その他	0.53	0.78	0.63	348
平均	0.61	0.45	0.47	1128

タに 1.0, ガンマ値に 0.1, カーネルに rbf を使用し, RF では決定木の数に 110, 各決定木に用いる説明変数の数に sqrt を使用することとした。

3.2 評価方法

分類器の性能評価には、5 分割交差検定を用いた。5 分割交差検定では、全データを 5 つに分割し、4 つを学習データ、1 つをテストデータとして学習する。その後テストデータを入れ替え

表 9 低頻度語, 高頻度語除去後の重要度が高い単語

単語	DF 値	重要度
絵	401	0.017857
娘	479	0.012161
息子	402	0.011117
ページ	390	0.010712
反応	148	0.009001
色	154	0.008882
購入	135	0.008617
くる	448	0.008313
言う	317	0.008100
喜ぶ	181	0.007957
私	416	0.006245
ん	382	0.005475

て 5 回学習した結果の平均を分類器の性能とする手法である。

性能評価の指標には, 正解と判断したデータについてどの程度合っているかを表す適合率, 正解のデータに対して正解と予測された割合を表す再現率, そして適合率と再現率の調和平均である F 値を用いた。さらに, RF では単語ごとにどの程度分類に影響しているかを表す重要度を得られる。重要度の高い単語を見ることで, 分類に用いられている単語が絵本の分類に適したものであるか確認した。

3.3 評価実験

SVM と RF による分類実験の結果を表 5, 表 6 に示す。適合率は SVM が 0.67 と RF が 0.68 となったが, 再現率と F 値については SVM がそれぞれ 0.42, 0.45, RF がそれぞれ 0.40, 0.44 となった。

3.4 低頻度語, 高頻度語の除去による分類性能の改善

レビュー文中には, 文書出現頻度 (Document frequency: 以下, DF 値と記述) の低い単語 (以下, 低頻度語と記述) 及び DF 値の高い単語 (以下, 高頻度語と記述) が含まれている。低頻出語は機械学習をする際にノイズとなる。また, 高頻出語には一般語 (e.g., いる, 読む) のような分類の決定に意味を持たない単語が含まれている。今回は単語の DF 値を用いてこれらの単語を除去することとし, DF 値が 1 及び 500 以上の単語を除去した。

重要度と DF 値の散布図を図 1 に示す。図の左下を見ると, 重要度と DF 値共に低い単語が多くあることが分かる。左下の部分がどのくらいあるか調べるため, DF 値の頻度分布を図 2 に示す。その結果, DF 値が 1 と 500 以上の単語が全単語中の 45.1% を占めていることが分かった。DF 値が 1 の単語には「称す」「速攻」「苦しむ」「小学生」「遠目」「上手い」「毛糸」「結論」などの単語が, DF 値が 500 以上の単語には「いる」「読む」「する」「絵本」「なる」「の」「*」「本」が含まれていた。DF 値が 1 の単語については最も重要度が大きい「譲り受ける」でさえもほぼ 0 に近いことが分かった。比べて DF 値が 500 以上の単語では「読む」や「絵本」など絵本のレビューに頻出すると考えられる単語の重要度が高くなっており, 分類に影響を及ぼしていることが分かった。これらの結果より, DF 値が 1 と 500 以上の単語を除いた 2,039 個の単語を使用し再度 SVM

と RF での分類を行った (表 7, 表 8 参照)。適合率は全ての名詞, 動詞, 形容詞を使用した時に比べて低下したが, 再現率は SVM が 0.42 から 0.45, RF が 0.40 から 0.45 へ向上し, F 値についても SVM が 0.45 から 0.48, RF が 0.44 から 0.47 へと若干向上した。

4. 考察

SVM と RF を用いて分類を行った結果, SVM が 0.45, RF が 0.44 とどちらもほぼ変わらない性能であることが分かった。低頻度語, 高頻度語の除去後に同様の分類を行った結果, SVM が 0.48, RF が 0.46 となり, 名詞, 動詞, 形容詞を使用した時より分類性能が向上した。本稿の目的である子どもの反応タイプについては, 低頻度語, 高頻度語除去後の SVM と RF の F 値がそれぞれ 0.76, 0.72 の精度で分類できていることが確認された。子どもの反応タイプに関連する単語 (e.g., 娘, 息子, 反応) や, 絵柄に対する言及タイプに関連する単語 (e.g., 絵, 色) の重要度が高くなっていったことから, 低頻度語, 高頻度語を除去することで分類に有効な単語の重要度が向上したと考えられる。

また今回, 3 人の評価者によりレビュー文の分類を行い, 3 人中 2 人以上の分類が一致したレビュー文を機械学習に使用した。一方で, 3 人とも分類が一致せず, 不要文として除いた文を表 10 に示す。不要文の中には, 一つの文に複数の分類を含むような文が多く含まれていた。例えば「音のリズムや, シンプルな絵, 0~2 歳児くらいにはピッタリの絵本です。」という文では着目する部分によって分類が異なる。「音のリズム」に着目すると絵本の内容に関する文のため物語に対する感想タイプとなり, 「シンプルな絵」に着目すると絵柄に対する言及タイプ, 「0~2 歳児くらいにはピッタリの絵本です。」に着目すると対象年齢に対して言及しているため書誌に対する言及タイプとなる。このような文に対しては, もっと短い文に区切り分類を行うか, 一つの文に対して複数の分類を付けるマルチラベル方式で分類を行うことで正確な分類が可能になると考えられる。

その他の中には読み聞かせをした時の子どもの年齢や読み聞かせを行った者の立場などを表す文や, 読み聞かせ方の工夫について記述された文が含まれていた。子どもの年齢や読み聞かせを行った者の立場については, 「絵本ナビ」から取得可能であるため, 今後それらの情報と分類結果を組み合わせることでより詳細な検索が可能なシステムが構築できると考えられる。読み聞かせ方の工夫についての文も抽出することで, 読み聞かせを初めて行う養育者や現状子どもの反応がいまいちな絵本があるという養育者などが読み方の工夫を行う事ができ, 有用な文と言えるため, 分類の追加を検討している。さらに, 「おすすめ」といった単語が含まれているような推薦情報についてもその他に含まれていたため, これも分類の追加を行い, より詳細な分類を行いたい。それ以外の文には, 絵本を知ったいきさつに関する文や, 絵本の状態を示す文も含まれていた。

5. 展望

本研究では絵本に対する子どもの選好に着目した類似絵本検

表 10 評価者間で分類が一致しなかった文の例

レビュー文	評価者 1	評価者 2	評価者 3
一冊でたくさんのお楽しみ方ができるともおすすめの一冊です。	1	その他	5
1歳代ではくだものや動物に反応し、単語を覚えるきっかけにもなりました。	1	その他	5
単純な言葉の繰り返しなので、乳児さんにはピッタリの絵本だと思います。	1	その他	3
「がたん、ごん…」の繰り返しの言葉のリズムとシンプルな絵が月齢の低い赤ちゃんにも理解できるようです。	1	4	3
意味がわからなくても、単純な言葉の繰り返しの絵本なのでリズムカルに読んでいるとすっかりお気に入りの絵本になりました。	1	その他	4
そして、しろくまちゃんの目がオレンジ色で二重になっているのが怖いです。	2	3	1
音のリズムや、シンプルな絵、0~2歳児くらいにはピッタリの絵本です。	2	3	1
絵本を通してやる気を出してくれると読んだ自身もうれしくなれます。	4	その他	1
子どもだけでも楽しめる絵本だと思います。	その他	4	1

索システムの構築を目指し、SVM と RF を用いた機械学習によって絵本のレビューを自動分類する判別器を作成した。その結果、SVM と RF 共に全体としての分類精度はそれぞれ 0.48, 0.47 と低かったが、本研究で特に収集したい子どもの反応タイプについてはどちらも 0.70 以上の精度で分類できることが分かった。さらに、低頻出語と高頻出語を除くことで全体の分類精度が向上することが確認できた。

今後はさらにデータを増やし、機械学習に用いる各分類の文の数をバランスすることで分類精度の改善を検討している。また、絵本のレビューに加え、出版社が公開している絵本のあらすじ情報を追加することで、誤分類の多かった物語内容に対する分類性能の向上も検討している。

さらに、分類することができたレビュー文をネガティブなレビュー文かポジティブなレビュー文かに自動分類することで、どのような選好の子どもがどのような絵本を好きか、または嫌いかの判別が可能となることが期待される。

6. おわりに

絵本に対する子どもの選好を考慮し絵本を検索することができるシステム構築の足がかりとして、本稿では絵本に対する子どもの選好の収集を試み、機械学習によって絵本のレビューを自動分類する判別器を作成した。その結果、子どもの反応に関する記述についてはどちらも高い精度で分類できること、低頻度語、高頻度語を除去することで全体の分類精度が向上することが確認された。

文 献

- [1] 全国出版協会・出版科学研究所, “2016 年版出版指標年報,” 2016.
- [2] 中西一弘, 覚道知津子, “子どもが選択する絵本と大人が選択する絵本に関する一考察: 幼児絵本におけるマンガの手法の問題を中心に,” 大阪教育大学紀要. V, 教科教育, vol.40, pp.223-238, 1992.
- [3] 松村 敦, 杉 七瀬, 宇陀則彦, “読み聞かせ時の反応に着目した絵本に対する子どもの好みの取得方法に関する検討,” 日本教育工学会論文誌, vol.32, no.Suppl., pp.125-128, 2008.
- [4] 湯川和秀, 森山政訓, 今城和宏, 上野康治, 金田重郎, “電子書籍を用いた興味度判定サービス提供手法,” 情報処理学会論文誌, vol.1, no.1, pp.12-21, 2011.
- [5] 今満亨崇, 松村 敦, 宇陀則彦, “子どもの様子に着目したお話のための絵本選択支援の研究,” 日本教育工学会論文誌, vol.35, no.Suppl., pp.109-112, 2011.
- [6] 藤田早苗, 服部正嗣, 小林哲生, 奥村優子, 青山一生, “絵本検

索システム「びたりえ」～子どもにぴったりの絵本を見つけます～,” 自然言語処理, vol.24, no.1, pp.49-73, 2017.

- [7] 松村 敦, 柿島大貴, 宇陀則彦, “子どもの質問に着目したソーシャル絵本推薦システム,” 情報知識学会誌, vol.21, no.2, pp.143-148, 2011.
- [8] H. Uehara, M. Baba, and T. Utsuro, “Detecting an infant’s developmental reactions in reviews on picture books,” 29th Pacific Asia Conference on Language, Information and Computation, pp.64-71, 2015.
- [9] 山澤美由起, 吉村宏樹, 増市 博, “Amazon レビュー文の有用性判別実験,” 情報処理学会研究報告, vol.2006-NL-173, no.53, pp.15-20, 2006.
- [10] 岩井秀成, 池田 郁, 土方嘉徳, 西田正吾, “レビュー文を対象としたあらすじ分類手法の提案,” 電子情報通信学会論文誌, vol.J96-D, no.5, pp.1222-1234, 2013.
- [11] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti, “Automatically assessing review helpfulness,” Proc. 2006 Conference on Empirical Methods in Natural Language Processing, pp.423-430, 2006.
- [12] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, pp.273-297, 1995.
- [13] A. Liaw and M. Wiener, “Classification and regression by randomforest,” R News, vol.2, no.3, pp.18-22, 2002.
- [14] 山西良典, 古田周史, 福本淳一, 西原陽子, “出現頻度と構文特徴を用いたレビュー構造の俯瞰のための評価視点の抽出,” 知能と情報, vol.27, no.1, pp.501-511, 2015.
- [15] 高野敦子, 池奥渉太, 北村泰彦, “因果関係に着目した口コミ情報からの評判情報抽出,” 人工知能学会論文誌, vol.24, no.3, pp.322-332, 2009.
- [16] V. Vapnik, The Nature of Statistical Learning Theory, Springer Science & Business Media, 2013.
- [17] L. Breiman, “Random forests,” Machine Learning, vol.45, pp.5-32, 2001.