3D4-4

絵本を基にした対象年齢推定方法の検討

Automatic Assessment of Target Age Based on Corpora of Picture Books and Textbooks

藤田 早苗[*1] 小林 哲生[*1] 平 博順[*1] 南 泰浩[*1] 田中 貴秋[*1]
Sanae Fujita Tessei Kobayashi Hironori Taira Yasuhiro Minami Takaaki Tanaka

[*1]NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

We aim to create gradual readability (or target age) measures from infants to elder children. For Japanese texts, several readability measures have been proposed, none of which is applied to texts for infants. Therefore, in this paper, we employ 190 picture books besides 94 textbooks of elementary school, then we investigate the applicability of two previous works to these texts. Both of the previous works show high performance, in addition, we achieve higher performance by combining them.

1. はじめに

絵本からの言語情報は、言語発達における幼児への重要なインプットであり、年齢や発達に応じた絵本の選定は重要だと考えられる。しかしながら、これまで行われてきた日本語を対象としたテキストの難易度測定方法や、読みやすさ (リーダビリティ) の測定方法の研究は、小学生以上から大人向けの文書 [2, 3, 4, 5] や、外国人日本語学習者を対象としたもの [7] が主であり、絵本を始めとする幼児を対象とした文の難易度 (対象年齢) を推定する手法の研究はなかった。そこで、本稿では、より小さな子供向け (以下、幼児) の文を含めた対象年齢の推定方法を検討する。

これまで、幼児を対象とするような文の対象年齢の推定方法が研究されてきていない一つの理由は、教科書のように明確な年齢推定の基準となるコーパスがなかったことだと考えられる。そこで、本稿では、比較的対象年齢が細かく設定されている絵本を、幼児を対象とした場合の基準コーパスとして利用することを提案する。

一方、小学生以上であれば、教科書を基準として利用した対象年齢推定の手法は複数提案されている [3, 5]。そこで、絵本と小学校の教科書の両方を基準コーパスとして、幼児から小学生にかけての対象年齢推定実験を行う。実験では先行研究で提案されている難易度測定手法を適用し、先行研究が幼児向けの文にも適用可能かどうかを調査、それぞれの結果を比較し、改良方法を探る。

2. 先行研究

教科書を基準とした難易度測定を行う研究には、佐藤ら [3, 4](§ 2.1) と、柴崎・玉岡 [5](§ 2.2。以下、柴崎方式) の研究がある。

2.1 帯

佐藤らは、小学校から大学までの教科書を用いて、13 段階の難易度を推定する難易度測定システム「帯」を構築し、公開している[*1]。難易度は、1 から 6 が、小学 1 年から 6 年、7 から 9 が中学 1 年から 3 年のように対応付けられる。難易度の規準には、小中高大の教科書 127 冊から抽出した 1478 サンプル、約 100 万字のコーパスを用いている。帯では、まず、それぞれの難易度に対する尤度を、連続する 2 文字の生起確率 (文字 bigram) に基づいて計算し、得られた尤度のうち、最大の尤度をとる難易度を求める難易度としている。また、難易度に順位関係が存在し、難易度に対する尤度は緩やかな曲線を描くことが期待されることから、尤度の値をクラス間でスムージングしている。帯では文字 bigram のみを用いるため、形態素解析や品詞体系に影響されないという利点がある。

連絡先: 藤田 早苗，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 ，Tel: 0774-93-5331，Fax: 0774-93-5345，fujita.sanae@lab.ntt.co.jp

∗1 http://kotoba.nuee.nagoya-u.ac.jp/sc/obi2/, 本稿では、obi ver. 2.304 を利用

2.2 柴崎方式

柴崎ら [5] は、小学 1 年から中学 3 年までの国語教科書中のテキストを基に、学年による文章の難易度の測定方法を提案している。利用したテキストは、国語の散文の教材のみ、243 テキスト、約 58 万文字、約 2 万文である。柴崎らは、複数の特徴量について調査した結果、テキスト全体のひらがなの割合と 1 文の平均述語数が、学年を推定するための有意な独立変数となることを示した。

柴崎らが導出した、学年を推定するための重回帰式は (1) の通りである。ただし、式 (1) において、Y = 学年、X_1 = テキスト全体のひらがなの割合、X_2 = 1 文の平均述語数 である。

$$Y = -0.145X_1 + 0.587X_2 + 14.016 \quad (1)$$

3. 実験データ

3.1 絵本データベース

本稿では、小学生以前の対象年齢を測るための基準コーパスとして、絵本データベース [6] のうち、対象年齢が細かく設定されている福音館書店の月刊誌 (以下、KODOMO)190 冊を用いる。KODOMO では、各絵本の対象年齢は 0・1・2 歳向け (以下、K012)、年少 (3 歳児) 向け (以下、K3)、年中 (4 歳児) 向け (以下、K4)、年長 (5 歳児) 向け (以下、K5) とわかれている。絵本データベースには、他にも、2010 年度の紀伊国屋書店グループの売上冊数が上位の絵本 1,010 冊が含まれるが、対象年齢が記載されていた絵本は、463 冊 (45.8 %) にとどまり、その記載方法も「3 歳から小学校初級むき」「乳児から」「4 才から」のように多様で、KODOMO のように 1 歳単位で対象年齢が設定されている絵本は少ない。

KODOMO の例を (1) に示す。例 (1) のように、絵本の本文データは、元のページのレイアウトに忠実に入力されており、

文や文節の途中での改行なども、そのまま再現されている。

また、KODOMO の全文に対し、IPA 品詞体系による形態素解析結果を人手修正した情報が付与されている [1]。

(1) まなちゃん、おばあちゃんの
おひざに　おいで　おいで
(長野ヒデ子・さく「まなちゃんのいす」p.2 (2011), 福音館書店, (K012))

3.2 BCCWJ

小学生の基準コーパスとしては、現代日本語書き言葉均衡コーパス[*2](以下、BCCWJ) に含まれる小学校の教科書を用いた。BCCWJ では、(2) のような xml 形式で配布されている。ここで、<sentence> タグの付いている部分を抽出し、文 (3) のようにタグを取り除いたものを利用した。

(2) <sentence> そんな <ruby rubyText="きょく"> 曲 </ruby> をきいてみましょう。</sentence>
(OT61_00052 : 畑中良輔 ほか著, 小学生の音楽 3 (2006), 教育芸術社)

(3) そんな曲をきいてみましょう。

表 1 に、絵本と教科書のデータサイズを示す。なお、BCCWJ には様々な教科の教科書が含まれており、内訳は、芸術 26, 数学 23, 国語 17, 理科 12, 社会 10, 技術家庭 3, 生活 3 だった。

表 1: 実験データのサイズ

	学年	クラス	数[*3]	行数 計	行数 平均	文字数 計	文字数 平均
	K012	1	27	494	18.3	3,903	144.6
絵	K3	2	30	1,221	40.7	14,141	471.4
本	K4	3	68	4,984	73.3	82,675	1215.8
	K5	4	65	7,064	108.7	118,942	1829.9
		小計	190	13,763	72.4	219,661	1156.1
	小 1	5	9	271	30.1	5,719	635.4
教	小 2	6	10	398	39.8	9,173	917.3
	小 3	7	19	1,087	57.2	27,757	1460.9
科	小 4	8	15	1,040	69.3	28,319	1887.9
	小 5	9	20	1,576	78.8	45,805	2290.3
書	小 6	10	21	2,483	118.2	68,963	3284
		小計	94	6,855	72.9	185,736	1975.9
		合計	284	20,618	72.6	405,397	1427.5

*3「絵本」の場合冊数、教科書の場合ファイル (テキスト) 数を示す

4. 先行研究の適用実験

本章では、前章で紹介した基準コーパスに対し、先行研究の手法を適用する。難易度のクラスと対象年齢は、表 1 の通り、012 歳児がクラス 1、3 歳児がクラス 2、のように対応する。

4.1 帯の適用

帯の配布プログラムには、規準コーパスから難易度推定モデルを作成する機能がついている。スムージング方法は選択できるが、本稿では、スムージングなし、2 次曲線スムージング、4 次曲線スムージングの 3 つの方法で得たクラスの中央値を用いた。

表 2 に、leave-one-out cross-validation を行った結果を示す。ここで、± 0 は正しいクラスを推定できた数と割合 (的中率)、± 1 は、前後 1 クラスにずれて推定された場合も正解とした場合の数と割合を示す。また、表の下部には、正解クラスと推定されたクラスの相関係数 (R)、及び、二乗平方根誤差 (Root Mean Square Error; $RMSE$) を示している。絵本の場合、先行研究で用いられたコーパスに比べ、非常に有効 bigram が少ない絵本が相当数ある。しかし、教科書を用いた場合 ($R = 0.94$, $RMSE = 1.207$)[3] と、同等以上の精度が得られることがわかった。

4.2 柴崎方式の適用

柴崎方式 [5] と同様に、テキスト全体のひらがなの割合 (X_1) と 1 文の平均述語数 (X_2) を独立変数とし、対象年齢のクラス (Y) を予測するための重回帰式を導出した。平均述語数を出すための形態素情報は、絵本の場合、人手修正されたデータを用いた。また、教科書データは、ひらがなの多いデータに対する形態素解析精度を向上させた形態素解析モデル [1] による自動解析結果を用いた。

ただし、柴崎らは学習データとして国語の散文のみを利用し、かつ、一度構築した重回帰式を用いて外れ値を検出し、外れ値となったテキストは利用せずに重回帰式を再導出しているが、本稿ではこれらの処理は行なっていない。同様の処理を行うと学習に利用できるテキストが非常に限られてしまうことと、帯と同じテキストでの評価ができなくなるためである。

表 3 に、leave-one-out cross validation の結果を示す。ただし、leave-one-out cross validation を行うため、1 ファイルをテストデータとして取り除いた状態で、係数の再導出とクラス推定を繰り返している。また、Y は離散値ではなく連続値として得られるため、小数第一位の四捨五入によっていずれかのクラスに振り分けている。つまり、$Y = 1.4$ と得られれば、クラスは 1 とした。さらに、$Y < 0.5$ の場合には、クラス 1 に、$Y >= 10.5$ の場合クラス 10 とした。

柴崎ら [5] は、決定係数は $R^2 = 0.791$ (調整済み $R^2 = 0.789$) と報告している。本稿のデータに適応した場合、$R^2 = 0.76$ (調整済み $R^2 = 0.759$) となり、教科書だけの場合ほどではないが、絵本を含めた場合でも高い予測精度であることがわかる。また、柴崎らの方法と同様に、外れ値を除くなどの処理を行えば、更に高い精度が得られる可能性もある。

なお、全コーパスを学習に用いて導出した重回帰式は、式 (2) の通りである[*4]。ただし、式 (2) において、Y = 難易度クラス、X_1 = テキスト全体のひらがなの割合、$X_2 = 1$ 文の平均述語数 である。

$$Y = -0.06086X_1 + 2.96780X_2 + 5.58687 \qquad (2)$$

4.3 比較と分析

本節では、帯を適用した場合 (表 2) と、柴崎方式を適用した場合 (表 3) を比較する。全体的な精度は、± 0 も ± 1 も帯の方が高い。しかし、比較的小さな子供のクラスの推定は柴崎方式の方が良い推定結果であることも多い。特に、クラス 1 (012 歳児) の場合、帯の 2 倍以上の的中率である。

柴崎方式では的中しているのに、帯だと 2 クラス以上離れたクラスだと推定されたものは、「だれかしら」[*5]「ほんやのおじさん」[*6]など、絵本ばかりだった。

逆に、帯では的中しているが、柴崎方式だと 3 クラス離れたクラスだと推定されたものは、小学 4 年以上の教科書に集中している。絵本の場合でも、帯では的中しているが、柴崎方式だと 2 クラス離れたクラスだと推定されるものもあり、絵本だからといって必ずしも柴崎方式の方が良いわけではないが、柴崎方式は、より幼い子供向けの文かどうかの判定に有利であ

*2 http://www.ninjal.ac.jp/kotonoha/

*4 いずれの変数も $p < 0.001$ で難易度クラスと有意に相関がある

*5 佐々木マキ さく「だれかしら」(2011), 福音館書店, (K012)

*6 ぶん・ねじめ正一, え・南 伸坊「ほんやのおじさん」(2011), 福音館書店 (K3)

り、帯はより大きな子供向けの文の判定に有利であるという傾向がある。

これは、両手法が推定に利用する特徴に由来すると考えられる。帯では、文字 bigram を利用しているが (§ 2.1)、小学校では学年ごとに学習する配当漢字が決まっているため、文字 bigram が非常に有効に働くのだと考えられる。しかし、本稿で利用した絵本はすべて幼児対象であり、漢字はほとんど出現しない。また、文字数自体も非常に少なく、K012 の場合有効 bigram 数は平均 89.3、最も少ないものでは 29 しかない[*7]。そのため、特に小さな幼児向けの絵本では文字 bigram だけでは推定が難しいのだと考えられる[*8]。

一方、柴崎方式では、ひらがなの割合と平均述語数を特徴として用いている。前述のように、絵本では漢字はほとんど出現しないが、述語の数は対象年齢によって大きく異なることが考えられる。例えば、有効 bigram が 29 しかない「けろけろぴょん」の場合、述語はでてこず、正しいクラス 1 が推定できている。

4.4 両手法の組み合わせ

前節 (§ 4.3) で述べたように、2 つの先行研究は全く異なる手法であり、異なる傾向がある。そこで、本節では、両方の結果を用いることで精度の向上をはかる。

帯の配布プログラムでは、難易度に対する尤度のスムージング方法が 4 通り[*9]用意されており、スムージングなしを加えて 5 通りの推定結果を得ることができる。

そこで、(1) 帯の全てのスムージング方法によって得られた難易度を平均し、さらに、(2) 柴崎方式による推定結果 Y と平均し、(3) 小数第 1 位を四捨五入して難易度クラスを推定する。ここで、Y は、重回帰式から得られた値そのものを利用する。

例えば、「とけいのあおくん」[*10] の場合、正解の難易度クラスは 2 だが、帯の各スムージング方法で推定された難易度クラスは、3, 3, 3, 3, 4, である。一方、柴崎方式で得られる結果は、$Y = 5.169482$ であり、クラス 5 と推定された。ここで、帯のスムージング結果の平均値 3.2 と、柴崎方式の 5.169482 の平均値 4.184741 から、最終的にクラス 4 と推定する。

表 4 に結果を示す。表 4 から、両方の特徴が均され、より精度 ($\pm 0, \pm 1$) も相関係数も高い結果を得られたことがわかる。

5. まとめと今後の課題

これまで、日本語を対象としたテキストの難易度測定の研究には、小学生以前の幼児を対象とした研究はなかった。そこで、本稿では、幼児を対象とした文を含めた対象年齢の推定方法を検討した。

対象年齢を測るための基準コーパスとして、比較的対象年齢が細かく設定されている絵本と、教科書を利用し、幼児から小学生にかけての対象年齢推定実験を行った。

実験では、小学生以上を対象としたテキストの難易度推定を行う先行研究である、帯 [3]、および、柴崎らの手法 [5] を適用し、これらの手法が幼児向けの文にも有効かどうかを調査した。その結果、いずれの手法も幼児向けの文を含めた対象年齢推定にも高い性能を発揮することがわかった。また、特徴を調べたところ、前者はより高学年の判断に有効であり、後者は幼児むけの絵本の判定に有効であることがわかった。

本稿ではさらに、両方の手法を組み合わせる方法を提案し、より高い対象年齢の推定精度を得ることができた。ただし、組み合わせることにより、全体的な精度は高くなったが、一部のクラスでは個々の手法よりも精度が低下する場合もあるため、今後はさらに、両手法の良い点をうまく取り入れた方法を検討したい。

参考文献

[1] 藤田 早苗, 平 博順, 小林 哲生, 田中 貴秋. “絵本のテキストを対象とした形態素解析”, 自然言語処理, Vol. 21, (3), 2014, (to appear).

[2] 伊藤 美咲姫, 佐藤 理史, 駒谷 和範. “難しい日本語文の自動検出のための基礎調査”, 言語処理学会第 19 回年次大会 (NLP-2013). pp.886–889, (2013).

[3] 小島 健輔, 佐藤 理史, 藤田 篤. “文字 bigram モデルを用いた日本語テキストの難易度推定”, 言語処理学会第 15 回年次大会 (NLP-2009). pp.897–900, (2009).

[4] 佐藤 理史. “均衡コーパスを規範とするテキスト難易度測定”. 情報処理学会論文誌, Vol. 52, (4), pp. 1777–1789, (2011).

[5] 柴崎 秀子, 玉岡 賀津雄. “国語科教科書を基にした小・中学校の文章難易度学年判定式の構築”, 日本教育工学会論文誌 (2010), Vol. 33 (4), pp.449-458.

[6] 平 博順, 藤田 早苗, 小林 哲生, (2012). 絵本テキストにおける高頻度語彙の分析 情報処理学会関西支部 支部大会, F-103.

[7] 李 在鎬. “大規模テストの読解問題作成過程へのコーパス利用の可能性”, 日本語教育学会論文誌 (2011), 148 号.

∗7 田村ゆうこ さく, 「けろけろぴょん」 (2010), 福音館書店, (K012)
∗8 [3] では、学習に有効 bigram が 250 以上のコーパスを用いているが、特に低年齢の絵本では、有効 bigram が 250 以上ある絵本はほとんどない
∗9 2 次から 5 次の曲線スムージング
∗10 エリザベス・ロバーツ さく, 灰島かり やく, 殿内真帆 え, 「とけいのあおくん」 (2009), 福音館書店, (K3)

表 2: 帯 2 によりモデルを再構築した場合の推定結果 (leave-one-out cross-validation)

クラス		1	2	3	4	5	6	7	8	9	10	計	± 0	(%)	± 1	(%)
	1	6	11	7	3	0	0	0	0	0	0	27	6	22.2	17	63.0
絵	2	2	7	12	9	0	0	0	0	0	0	30	7	23.3	21	**70.0**
	3	3	7	46	12	0	0	0	0	0	0	68	46	**67.6**	65	**95.6**
本	4	1	1	26	31	6	0	0	0	0	0	65	31	47.7	63	**96.9**
(小計)												(190)	(90)	(47.4)	(166)	(87.4)
	5	0	0	0	0	6	3	0	0	0	0	9	6	66.7	9	100.0
教	6	0	0	0	0	4	3	2	1	0	0	10	3	30.0	9	90.0
	7	0	0	0	0	0	0	7	11	1	0	19	7	36.8	18	94.7
科	8	0	0	0	0	0	0	2	5	8	0	15	5	33.3	15	100.0
	9	0	0	0	0	0	0	0	3	15	2	20	15	75.0	20	100.0
書	10	0	0	0	0	0	0	1	0	13	7	21	7	33.3	20	95.2
合計		12	26	91	55	16	6	12	20	37	9	284	133	46.8	257	90.5

$R = 0.94$, $RMSE = 0.951$

表 3: 柴崎方式と同様にモデルを再構築した場合の結果 (leave-one-out cross-validation)

クラス		1	2	3	4	5	6	7	8	9	10	計	± 0	(%)	± 1	(%)
	1	13	8	5	1	0	0	0	0	0	0	27	13	**48.1**	21	**77.8**
絵	2	2	8	9	10	1	0	0	0	0	0	30	8	**26.7**	19	63.3
	3	3	8	19	29	8	1	0	0	0	0	68	19	27.9	56	82.4
本	4	3	6	7	38	11	0	0	0	0	0	65	38	**58.5**	56	86.2
(小計)												(190)	(78)	(41.1)	(152)	(80.0)
	5	0	0	0	1	6	0	1	1	0	0	9	6	66.7	7	77.8
教	6	0	0	0	0	4	3	3	0	0	0	10	3	30.0	10	100.0
	7	0	0	0	0	2	3	8	2	3	1	19	8	42.1	13	68.4
科	8	0	0	0	0	1	1	4	5	2	2	15	5	33.3	11	73.3
	9	0	0	0	0	0	3	2	2	8	5	20	8	40.0	15	75.0
書	10	0	0	0	0	0	0	7	7	4	3	21	3	14.3	7	33.3
合計		21	30	40	79	33	11	25	17	17	11	284	111	39.1	215	75.7

$R^2 = 0.76$, 調整済み $R^2 = 0.759$, $R = 0.87$, $RMSE = 1.309$

表 4: 組み合わせ: 帯の複数のスムージング方式による結果の平均値と柴崎方式の平均値 (leave-one-out cross-validation)

クラス		1	2	3	4	5	6	7	8	9	10	計	± 0	(%)	± 1	(%)
	1	9	11	6	1	0	0	0	0	0	0	27	9	33.3	20	74.1
絵	2	0	9	14	7	0	0	0	0	0	0	30	9	30.0	23	76.7
	3	0	10	33	23	2	0	0	0	0	0	68	33	48.5	66	97.1
本	4	3	0	16	44	2	0	0	0	0	0	65	44	67.7	62	95.4
(小計)												(190)	(95)	(50.0)	(171)	(90.0)
	5	0	0	0	0	7	1	1	0	0	0	9	7	77.8	8	88.9
教	6	0	0	0	0	4	4	2	0	0	0	10	4	40.0	10	100.0
	7	0	0	0	0	0	2	10	4	3	0	19	10	52.6	16	84.2
科	8	0	0	0	0	0	1	2	8	3	1	15	8	53.3	13	86.7
	9	0	0	0	0	0	0	2	4	10	4	20	10	50.0	18	90.0
書	10	0	0	0	0	0	0	1	10	7	3	21	3	14.3	10	47.6
合計		12	30	69	75	15	8	18	26	23	8	284	137	48.2	246	86.6

$R = 0.94$, $RMSE = 0.937$