4Q-06

子供 Web コーパス作成に向けた子供向けページ判別法

佐藤 倫太郎† 安藤 一秋‡
香川大学大学院工学研究科† 香川大学創造工学部‡

1. はじめに

近年，小学校から高等学校までの教育機関を中心に，新聞を活用する教育（NIE: Newspaper in Education）が実施されている．しかし，新聞記事に出現する語句は子供にとって難しい場合が多く，小学校での NIE では，学習者が正しく記事内容を理解できない問題がある．この問題を解決するため，新聞記事に出現する難しい語句を平易に言い換える手法[1]が研究されているが，70%の正解率に留まる．

新聞記事に現れる語句を言い換えるためには，言い換え知識が必要である．子供を対象とした既存の言い換え知識として小学国語辞典があるが，語彙数が少ない問題がある．そのため，言い換え知識を新たに獲得するための情報源が必要である．言い換え知識を獲得するための情報源として，コーパスの利用が考えられるが，子供向けのテキストを十分に収集したコーパスは存在しない．

そこで，本研究では Web 上の子供向けテキストを大量に収集することで「子供 Web コーパス」を構築し，当該コーパスから言い換え知識を獲得することを目指す．膨大な Web ページから，子供向けページを効率よく収集するには，子供向けページを判定する手法が必要である．本稿では，SVM（Support Vector Machine）を用いた子供向けページの判定法と，1 文単位における子供向けセンテンス判定法を提案する．

2. 関連研究・先行研究

テキストの平易化に関する研究としては，梶原の研究[2]がある．梶原は，English Wikipedia から単言語パラレルコーパスを構築し，単語分散表現から導かれる文間類似度によって難解な文と平易な文のアライメントを求めている．外部知識に依存しない手法であるが，日本語による評価は行われておらず，コーパス構築や子供向けのテキスト判定に直接的に関連する研究ではない．

我々の先行研究である泉川の研究[3]は，広範囲に子供向けページを収集する方法として，子供向けポータルサイト内のリンクから収集する方法や，子供向けサイトのトップページより，その内部ページを子供向けとして収集する方法などを提案した．しかし，いずれも精度が低い結果となった．また，その結果から，泉川はページ単位の判定手法として，SVM による判定の可能性を示唆しているが，素性の具体的な検討や，判定の実現には至っていない．

3. 子供向け Web ページの判定法

先行研究において未実現であった，SVM を利用した子供向けページ判定法を提案する．

A Method to Distinguish Kids' Pages from the Web for Building Web Corpus for Kids
Rintaro Sato†, Kazuaki Ando‡
† Graduate School of Engineering, Kagawa University
‡ Faculty of Engineering and Design, Kagawa University

3.1 コーパス・モデルの構築

SVM に与える学習データは，その用途を Web ページ判定とすることから，Web 上から収集することが好ましい．しかし，事前調査[4]の結果，客観性が担保された膨大な子供向けあるいは一般向けテキスト群を準備することは困難であることを確認した．よって，本稿では，学習データとして，その対象読者が明らかなコーパス群を活用する．子供向けテキストとして，NHK が運営する子供向けニュースサイト「NEWS WEB EASY（以下，NWE）」の記事 300 件，および，小学校教科書コーパス（国語，社会について 4 年，5 年，6 年分：計 1,493 ページ相当）を利用する．一般向けテキストとして，読売新聞の 44,692 記事，神戸新聞の 2,823 記事を利用する．これらを組み合わせ，表 1 に示す 5 つのモデル（コーパス・モデル群）を利用する．

表 1 コーパス・モデル

モデル	子供向け	一般向け	件数
A	教科書コーパス	神戸新聞	1,493
B	教科書コーパス	読売新聞	1,493
C	NEWS WEB EASY	神戸新聞	300
D	NEWS WEB EASY	読売新聞	300
E	教科書コーパス＋ NEWS WEB EASY	神戸新聞	796

3.2 SVM に与える素性

本稿では，先行研究[3,4]，岩田らの研究[5]，やさ日チェッカーα版[6]，テキストの読み易さ解析ツール「TRF」[7]を参考に選定した以下の 7 素性を組み合わせて，SVM での学習に利用する．

(1) 難易度推定システム「帯 2」の推定難易度（obi2）
(2) テキスト内での漢字の占める割合（漢字割合）
(3) 動詞のみについて異なり語の割合（異なり）
(4) jReadability（jred）
(5) 係り受け木の深さの平均（係受平均）
(6) 係り受け木の深さの分散（係受分散）
(7) 日本語教育語彙表の難易度別語彙割合（語彙表）

(1)の obi2 は，小島らの開発した日本語テキスト難易度判定システム[8]の出力である．(4)の jReadability とは，李らが開発した日本語学習者における文章難易度判別システム[9]の出力値である．(7)の語彙表は，6 段階の難易度に分類された約 18,000 語の日本語教育用語彙を収録した「日本語教育語彙表」[10]を参照し，各難易度の語彙がテキスト中に占める割合である．(1)から(6)の各素性はそれぞれ 1 次元，(7)は 7 次元で表現する．

3.3 子供向け Web ページ判定法の性能評価

3.2 節で示した素性を組み合わせて，未知のデータを判定することにより，子供向けページ判定法の性能を評価する．学習データには，各コーパス・モデルを，評価データには，人手により収集した子供向け Web ページ，一般向け Web ページ，各 20 件を使用する．

まず，個々の素性単体のみを与えて判定し，F 値を調査する．各モデルにおける結果を表 2 に示す．表 2 に示す結果から，(1)obi2，(2)漢字割合，(7)語彙表が，全てのモデルに対して有効であり，また，(2)漢字割合は教科書を含むモデルに，(1)obi2 は NWE を含むモデルに特に有効であることが確認された．

この結果より，(3)異なり語以外の素性について，有効と思われる組み合わせに限定して実験を行った．そのうち，最も優れていた素性の組み合わせ（(1)obi2 と (2)漢字割合, (4)jred, (6)係受分散, (7)語彙表）の結果（F 値）を表 3 に示す．表 3 に示すとおり，全てのモデルにおいて F 値 0.85 を超える判定性能が確認された．なお，表 3 の「神戸・NWE」および「読売・NWE」は，(1)obi2 のみを素性に与えた際の結果よりも低い値となった．しかし，「NWE と教科書の混合・神戸」という，Web 上の多様なページ群に，より近いモデルにおいて，0.88 という結果を得たことから，総合的にこの素性組み合わせが適切であると考えられる．

表 2　素性単体での判定結果

		素性						
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
モデル	A	0.84	0.82	0.09	0.63	0.67	0.67	0.86
	B	0.84	0.83	0.08	0.67	0.64	0.67	0.88
	C	0.88	0.80	0.65	0.71	0.68	0.52	0.88
	D	0.88	0.80	0.67	0.79	0.68	0.67	0.88
	E	0.84	0.80	0.08	0.65	0.69	0.65	0.79
	平均	0.86	0.81	0.31	0.69	0.67	0.63	0.86

表 3　最適な素性の組み合わせによる結果（F 値）

神戸・教科書	読売・教科書
0.93	0.93
神戸・NWE	読売・NWE
0.86	0.86
NWEと教科書の混合・神戸	
0.88	

4. 1 文単位における子供向けセンテンス判定

Web 上には，子供向け・一般向けと一概に 2 分できないページも多く存在する[4]ため，ページ単位の判定の後，さらに 1 文単位でテキストを精査する必要がある．そこで，SVM を用いた子供向けセンテンスの判定法を検討する．

4.1 1 文単位判定モデル

SVM に与える学習データには，コーパス・モデルで使用したコーパスを文単位に分割したものを利用する．その詳細を表 4 に示す．

表 4　文単位のコーパス・モデル

モデル	子供向け	一般向け	件数(文)
A	教科書コーパス	神戸新聞	3,820
B	教科書コーパス	読売新聞	10,069
C	NEWS WEB EASY	神戸新聞	2,148
D	NEWS WEB EASY	読売新聞	2,148
E	教科書コーパス＋ NEWS WEB EASY	神戸新聞	4,296

判定には，3.2 節で示した「漢字割合」「語彙表」に「係り受けの深さ」を加えた 3 素性を利用する．

コーパス・モデルにおける学習データは，Web 上のページとは性質を異にし，全体で一貫して特定の読者層を意識して書かれている．したがって，コーパス・モデルについて，子供向け・一般向け学習データに含まれる各文は，全て子供向け・一般向けであるとみなし，10 分割交差検証で性能を評価する．評価結果の F 値を表 5 に示す．

何れのモデルにおいても，0.8 を超える性能が確認されたが，ページ単位の判定性能と比較すると，劣る結果となった．この背景には，学習データのうち，一般向けのデータ（神戸新聞・読売新聞）については，平易な文も存在することが影響していると考えられる．また，素性についてもコーパス・モデル程，検討できておらず，これらの点において，1 文単位の判定性能にはまだ改善の余地がある．

表 5　10 分割交差検証の結果（F 値）

神戸・教科書	読売・教科書
0.80	0.82
神戸・NWE	読売・NWE
0.84	0.86
NWEと教科書の混合・神戸	
0.83	

5. おわりに

本稿では，子供 Web コーパス構築に向けて，Web から子供向けテキストを収集するために必要な Web ページ単位および 1 文単位での子供向けテキスト判定法を提案した．

はじめに SVM を用いた子供向けページ判定法を提案した．次に，提案手法を用いて未知データを判定し，最高で 0.93 の F 値を確認した．次に，1 文単位の子供向けセンテンス判定において，モデルを構築し，10 分割交差検証によって判定性能を評価した結果，0.86 の F 値を得た．

今後は，1 文単位の判定について，学習データの改良および未知データで評価し，Web 上からクローリングしたページを判定することで，子供 Web コーパスを構築する．

謝辞

本研究の一部は，JSPS 科研費 16K00478 の助成を受けて実施した．本研究では，科研費（課題番号 25370573）の成果物である「日本語文章難易度判別システム」（http://jreadability.net）と，日本語学習辞書支援グループ (2015)「日本語教育語彙表 Ver 2.72b」(http://jisho.jpn.org/) を利用した．

参考文献

[1] 梶原他, “語釈文を用いた小学生のための語彙平易化”, IPSJ 論文誌, Vol56, No.3, pp.983-992, (2015).
[2] 梶原他, “平易なコーパスを用いないテキスト平易化のための単言語パラレルコーパスの構築”, IPSJ 第 229 回自然言語処理研究会, Vol.2016-NL-229, No.13, pp.1-8, (2016).
[3] 泉川他, “子供 Web コーパス構築のための子供向けページ判定手法の検討”, NLP 第 22 回年次大会論文集, pp.170-171, (2016).
[4] 佐藤他, “子供 Web コーパス構築のための子供向けページ判定法”, IPSJ 第 80 回全国大会講演論文集, No1, pp445-446, (2018).
[5] 岩田他, “子供による Web 検索のための検索結果リランク手法”, IPSJ 論文誌, Vol52, No.3, pp.1055-1068, (2011).
[6] やさ日チェッカーα版 Ver0.26 (http://www4414uj.sakura.ne.jp/Yasanichi1/checker/)
[7] 渡邉他, “TRF: テキストの読みやすさ解析ツール”, NLP 第 23 回年次大会発表論文集, pp.477-480, (2017).
[8] 小島他, “文字 bigram モデルを用いた日本語テキストの難易度推定”, NLP 第 15 回年次大会論文集, pp.897-900, (2009).
[9] 李在鎬, “日本語教育のための文章難易度研究”, 早稲田日本語教育学, Vol.21, pp.1-16, (2016).
[10] 日本語教育語彙表 Ver.2.72b (http://jisho.jpn.org/)