5L-09

子供 Web コーパス構築のための子供向けページ判定法

佐藤 倫太郎† 泉川 洸一郎† 安藤 一秋‡
香川大学大学院工学研究科† 香川大学工学部‡

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いため，小学校での NIE においては，学習者が正しく記事内容を理解できない問題がある．子供を対象とした新聞サービスも存在するが，一般新聞と比較して記事数や購読者数が少ないため，実践現場ではほとんど利用されていない．このような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研究[1]が進められている．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を提案し，実際にクローリングしてきたページ群を判定し，分類性能を評価する．

2. 関連研究・先行研究

子供コーパスの構築に関する研究として，坂本の研究[2]がある．坂本は，全国 4,950 校の小学校の Web サイトから小学生が書いた作文テキストを収集し，作文コーパスを構築している．このコーパスの収録語数は 123 万語を超えているが，子供が書いた作文から抽出されるテキストは，一般に感想から成る主観的なテキストであると考えられる．このようなテキストは，客観的な事実や概念を記す新聞記事の言い換え知識を抽出するための情報源に適していると言い難い．

テキストの平易化に関する研究としては，梶原の研究[3]がある．梶原は，English Wikipedia のみ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文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，子供向けのテキストを対象とした研究ではない．また，平易な文によるコーパスの構築も行っていない．

平易な文章に関する研究では，渡邉らの研究[4]がある．渡邉らは，入力されたテキストに対して，平均文長や語彙の難易度，容認度等の指標のスコアを計算し，出力するツール「TRF」を提案している．

我々の先行研究である泉川の研究[5]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子供向けページを取得する方法や，サイトのトップページのみを難易度推定システム「帯 2」[5]を用いて判定し，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ている．また，これらの結果から，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，その素性の具体的な検討や，分類の実現には至っていない．

3. SVM による子供向けページ判定法

先行研究で実現に至らなかった SVM を用いた子供向けページ判定手法について，具体的に素性を検討した[6]．先行研究および岩田らの研究[7]が注目した子供向けページの特徴を参考に検討した結果，以下の 6 素性を採用することにした．

1. 難易度推定システム「帯 2」の推定難易度
2. HTML 内のルビタグの有無（ふりがな有無）
3. テキスト内での漢字の占める割合
4. 平易な文末表現の割合
5. 括弧内ひらがな文字列の有無
6. 異なり語の割合（動詞のみ）

これらの素性を組み合わせ，10 分割交差検証によって SVM の分類性能を評価した．そのうち最も F 値の高かった，全ての素性を与えて子供向けページを判定した結果を表 1 に示す．

表 1 全ての素性を用いた判定（提案手法）の結果

素性	適合率	再現率	F 値
改善素性	0.96	0.94	0.95

表 1 より，適合率と再現率が共に高く，提案手法が高い分類性能をもつことを確認した．しかし，このモデルでは学習データが子供向け，一般向けともに 200 件と乏しい．そこで，膨大な未知の Web ページに対しても，上記の分類性能が得られるかどうかを評価する必要がある．

次節では，当該モデルによって Web 上からクローリングしてきた未知データから子供向けページを判定することで，分類性能を評価する．

4. SVM による子供向けページ判定法の評価

本節では，提案手法の分類性能を評価するため，クローリングしてきた膨大な Web ページ群に対して子供向けページ判定を行い，その結果を人手で評価する．

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

4.1 評価データの収集

評価データは，クローリングによって収集する．Web 上に存在するページには，一般向けページと比較して子供向けページは僅少であると考えられる．よって，より子供向けページを含むページ群を収集するため，クローラーに与えるシードは，「キッズ@nifty」のリンク集から得られた 688 件を使用する．これら各々のリンクに対して，その内部ページを全てクローリングした．そのうち，本文を抽出できなかったページを除外した結果，140,573 件が獲得できた．これを評価データに利用する．

4.2 SVM による判定

4.1 項で収集した評価データに対して，3 節で述べたモデルを用いて，子供向けページ判定を行う．SVM は，scikit-learn の LinerSVC 使用する．カーネルは線形カーネルであり，そのパラメータは既定値を用いた．

実験の結果，29,113 件のページが子供向けであると判定された．このうち，異なるシードからクローリングされたページ 100 件をランダムに抽出し，人手によって「子供向け」「グレー」「一般向け」「ノイズ」の 4 つに分類した．「グレー」は，部分的に子供向けではあるが，総合的に子供向けと判定するには問題があるページを，「ノイズ」は，メニュー項目等が混入した，コーパス構築には不適切なページを示す．その結果を表 2 に示す．

表 2 人手による評価結果

子供	グレー	一般	ノイズ
36	12	41	11

子供向けのページを正しく判定できた件数は 100 件中 36 件であり，グレーを含めても 48 件となった．これは，事前に行った 10 分割交差検証の結果よりも著しく低い結果である．また，子供向け，一般向けの何れにも分類できないノイズが含まれていることも確認された．

次節では，分類結果について分析・考察し，その要因や，今後の改善点について述べる．

5. 判定結果の分析・考察

5.1 一般向けページが誤判定された要因

子供向けと判定された一般向けページ群は，大きく分けて以下の 2 種類のパターンに分類できた．

(1) よみがなを含むが，難解な語句を使用している

(2) ブログ的な文体であり，語彙等が子供向けではない

(1)は，さらに 2 種に分類できる．1 つは，子供向けに書こうという意図はあるものの，難解と思われる語句によみがなを施すだけであり，その他の読みやすい工夫が不十分な場合である．一方は，一般向けに書かれた文章であり，大人でも読解が困難と思われる語彙によみがなが施されている場合である．特に前者の場合は，ある程度テキストの難易度を下げる効果がある，最も簡易的な手法である「よみがな」に，ページ作成者が頼り過ぎることが原因であり，素性として単によみがなの有無を与えることの問題点を示唆している．

(2)に該当するページは，語りかけるようなブログ的な文章（イベントのレポートや，映画やドラマの紹介など）であり，平易な文末表現や語彙も見られるが，随所に難解な語句が見られるようなページである．SVM に与えた素性に，「平易な文末表現の割合」がある．これは学習データの子供向けページに多く見られた「よ，ますか，かな」といった文末表現の割合であるが，ブログ的な文章には，一般向けであってもこのような表現が用いられるため，誤判定につながったと考えられる．

なお，グレーに該当するページ群は，一般向けに誤判定されたページ群と同様に，部分的に難解な語彙・表現を含むものの，一般向けと断定するにはやや易しいと思われたページ群である．

5.2 SVM に与える素性の問題点

5.1 項で示した，誤判定に繋がった要因のうち，(1)の要因から，素性として単に「よみがなの有無」を与えることが安直であることが示唆された．これについては，一定難易度以下の語彙に対するふりがなに限定するといった工夫が必要である．また，子供向けページに見られる「平易な文末表現の割合」についても，一般向けに作成されたブログ様のコンテンツにも同様に平易な文末表現が用いられるため，再考が必要である．

6. おわりに

本稿では，子供 Web コーパス構築に向けて，Web 上のから子供向けテキストを収集するために必要となる，子供向けページの判定手法を提案した．また，提案手法について評価を行い，その結果から問題点を考察した．

まず，SVM を用いた子供向けページ判定手法について，その素性を中心に簡潔に説明した．次に評価データの収集法について述べ，実際に収集してきたページ群から 100 件を抜粋し，子供向けページ判定を行った結果を示した．最後に，その結果から提案手法の素性を考察し，幾つかの素性は子供向けページ判定に相応しくないことが確認された．

今後は，素性と学習データの再構築し，判定法を改良する．判定法を確立した後は，Web 上からクローリングしたページを順次判定し，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．

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”, 情報処理学会論文誌, Vol56, No.3, pp.983-992, (2015).
[2] 坂本, “小学生の作文コーパスの収集とその応用の可能性”, 自然言語処理, Vol.17, No.5, pp.75-98, (2010).
[3] 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”, 情報処理学会第 229 回自然言語処理研究会, Vol.2016-NL-229, No.13, pp.1-8, (2016).
[4] 渡邉他, "TRF: テキストの読みやすさ解析ツール" , 言語処理学会 第 23 回年次大会 発表論文集, pp477-480, (2017)
[5] 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”, 言語処理学会第 22 回年次大会論文集, pp.170-171, (2016).
[6] 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の検討”, FIT2017 第 16 回情報科学技術フォーラム講演論文集, 第 2 分冊, pp117-118, (2017)
[7] 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”, 情報処理学会論文誌, Vol52, No.3, pp.1055-1068, (2011).