D-002

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
A Method to Distinguish Kids' Pages from the Web for Construction of Web Corpus for Kids

佐藤 倫太郎† 泉川 洸一郎† 安藤 一秋‡
Rintaro Sato Koichiro Izumikawa Kazuaki Ando

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手法について検討する．

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．

我々の先行研究である泉川の研究[4]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．

† 香川大学大学院工学研究科 Graduate School of Engineering, Kagawa University
‡ 香川大学工学部 Faculty of Engineering, Kagawa University

3. SVM による子供向けページ判定法の検討

本稿では，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，学習データの収集および各種素性について検討し，評価する．さらに，その結果を踏まえ，素性の改善を検討し，同条件で評価する．

3.1 SVM に与える学習データの構築

SVM に与える学習データとして，子供向けテキストと一般向けテキストが必要になる．このうち，子供向けテキストは，子供向けサイトのトップページにあたる URL をシードとして，そのページ以下をクローリングすることで子供向けページを収集し，それらのページに記載されているテキストを利用する．

シードの収集法について述べる．まず「Yahoo!きっず」のリンク集から，明らかに子供向けと判断できるサイトを人手で選出し，そのトップページをシードとして収集する．さらに，一般向けサイト内のコンテンツの一部として子供向けコンテンツが含まれる場合がある．こうした子供向けページも収集するため，人手で地方公共団体等のサイトにアクセスし，子供向けコンテンツのトップページもシードとして収集する．

学習データのうち，一般向けテキストは，人手によって一般向けと判断されたサイトを広く選択し，そのトップページをシードとして，それ以下のページをクローリングすることで一般向けページを収集し，それらのページから抽出したものを利用する．

上記の方針にしたがって子供向けサイトのシード 85 件，一般向けサイトのシード 90 件を人手で選出し，クローリングを行った．続いて，収集されたページから文字数が 0 となるページを除外した結果，一般向けとして 2,001 ページ，子供向けとして 2,863 ページが収集できた．その後，一般向け及び子供向けページから，さらに人手でそれぞれ 200 ページずつを抜粋し，合計 400 ページに記載されているテキストにラベルを付与し，学習データとして利用する．

3.2 SVM による判定法の性能評価

SVM の分類性能は，10 分割交差検証で評価する．SVM に与える素性として，泉川の研究および岩田らの研究[6]が注目した子供向けページの特徴を参考に，以下の 4 つを用いる．

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

これらの素性を組み合わせ，全 14 通りについて実験を行った．そのうち最も F 値の高かった，全ての素性（以下，提案素性）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 提案素性による判定結果

適合率	再現率	F 値
0.94	0.89	0.92

表 1 より，再現率が適合率を下回っていることがわかる．この点から，子供向けページが一般向けと誤判定される割合は低いが，子供向けページの取りこぼしに問題があるといえる．そこで，一般向けと誤判定された子供向けページを調べたところ，2 の素性について，ふりがながルビタグではなく，漢字に後続する括弧内に記載されている場合が散見された．また，提案素性が捉える特徴を呈さない子供向けに書かれたページが存在することも確認された．よって，2 の素性の改善及び新たな素性を追加することで，判定性能を向上させる必要がある．

3.3 素性の改善

先の実験のエラー分析を基に「2. HTML 内のルビタグの有無」では取りこぼしてしまう，ルビタグを用いないふりがながあるページを拾うために「括弧内ひらがな文字列の有無（括弧かな有無）」を導入する．また，新しい素性として，やさしい日本語を判定するサービス「やさ日チェッカーα版 Ver0.25b[7]」の診断基準を参考に「異なり語の割合（動詞のみ）」を導入する．

表 2 に，二つの新規素性をそれぞれ単一で与えた場合の判定結果を示す．なお，実験環境は全て 3.2 の環境と同じである．

表 2 新規素性のみを用いた判定結果

素性	適合率	再現率	F 値
括弧かな有無	0.97	0.87	0.92
異なり語の割合（動詞のみ）	0.52	0.55	0.53

素性「括弧かな有無」については，単体のみでも高い F 値を示した．適合率が高い一方，ふりがなが振られない子供向けページも存在することから，再現率は比較的低い．「異なり語の割合（動詞のみ）」は F 値が 0.53 と低い．しかし，異なり語の割合はページの文字数にも影響を受けると考えられるため，今後，評価データセットを変えて再評価する必要がある．

次に，二つの新規素性と提案素性を組み合わせた素性（以降，改善素性）を用いた場合の判定結果を表 3 に示す．

表 3 改善素性を用いた判定結果

素性	適合率	再現率	F 値
提案素性	0.94	0.89	0.92
改善素性	0.96	0.94	0.95

提案素性と比較し，F 値が 0.03 向上し，特に改善すべき再現率が 0.05 ポイント向上した．取りこぼしが大幅に改善された一方で，適合率の低下は見られず，これも 0.02 ポイント向上している．依然として適合率よりも再現率が低いが，Web 上に存在する膨大なページから子供向けページを収集するタスクにおいては，本再現率により十分な子供向けページの収集が期待できるといえる．今後は，質の高いコーパスの構築のため，再現率を維持しつつ，適合率の向上に注力する必要がある．

4. おわりに

本稿では，子供 Web コーパスの構築に向けて，Web 上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について検討した．

まず，SVM の学習データの構築，および 4 つの素性を提案した．その後，提案素性について判定性能を評価した結果，再現率が低く，コーパス構築のためには性能が不足していることがわかった．また，素性「ふりがなの有無」に問題があることに加え，新たな観点からの素性が必要であることを確認した．次に，新たな素性として「HTML 内のルビタグの有無」および「異なり語の割合（動詞のみ）」を追加することによって，SVM の F 値が向上することを確認した．特に再現率については，0.05 ポイントと大きく改善され，子供向けページの取りこぼしが減少した．

今後は，より質の高いコーパス構築のため，再現率の維持と適合率の向上を図る．また，素性値の正規化や，異なる評価データセットでの実験等も行い，最良の判定性能を得る．その後，その分類器を以て Web 上からクローリングしたページを順次判定して子供 Web コーパスを構築し，言い換え知識の獲得を目指す．

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] 梶原智之, 山本和英, “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] 坂本真樹, “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] 梶原智之, 小町守, “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（第 3 回自然言語処理シンポジウム）, Vol.2016-NL-229, No.13, pp.1-8, (2016).
[4] 泉川洸一郎, 安藤一秋, “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).
[5] 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 回年次大会論文集, pp.897-900, (2009).
[6] 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).
[7] やさ日チェッカーα版 Ver0.25b http://www4414uj.sakura.ne.jp/Yasanichi1/checker/