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

泉川 洸一郎 [1]　　安藤 一秋 [2]

[1] 香川大学大学院工学研究科　　[2] 香川大学工学部

[1] s14g453@stmail.eng.kagawa-u.ac.jp　　[2] ando@eng.kagawa-u.ac.jp

1. 研究背景

近年，小学校などの教育機関では，新聞を教材として活用する教育（NIE : Newspaper in Education）が実施されている．NIE の実態調査結果報告書[1]によると，NIE を実践することで，児童の読解力と表現力が向上することや，社会に対する関心が高まることなどが報告されている．しかし，新聞に出現する語句は，小学生にとって難しいものが多く，記事の理解は容易ではない．

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小学生が理解できる平易な語句に言い換える研究[2]が進められている．難しい語句を平易に言い換えるためには，言い換え知識が必要となる．小学生を対象とした言い換え知識として小学国語辞典が利用できるが語彙数が少ない．したがって，語彙数問題を解決するためには，他の情報源から言い換え知識の自動獲得が必要となる．

近年，大規模コーパスが整備されており，様々な用途に活用できるようになった．しかし，子供向けに書かれたテキストを大量に集めたコーパスは，現在存在していない．

本研究では，Web 上の子供向けサイトに存在する平易な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，コーパスから言い換え知識を自動獲得する手法の実現を目的とする．本稿では，クローリングにより子供向けサイトを効率良く収集するために，子供向けページの判定手法を検討する．

2. 子供向けの換言に関する関連研究

梶原らは，学習基本語彙ではない語を難解語と定義し，難解語を学習基本語彙に言い換える手法[2]を提案している．複数の国語辞典から難解語を見出し語として検索し，定義文から難解語と同一品詞の学習基本語彙を換言先候補として取り出す．そして，取り出した換言先候補を語の類似度など複数の指標を用いて言い換える．梶原らは言い換え知識を取得する国語辞典の一つとして小学国語辞典を使用しているが，収録されている語彙数が少ないことが問題であると指摘している．

本稿では，Web 上の子供向けサイトから平易なテキストを収集して子供 Web コーパスを構築し，コーパスから言い換え知識を収集することで，この問題の解決を目指す．

3. 子供向けサイトの収集手法の検討

Web 上から子供向けに書かれたテキストを収集するため，子供向けポータルサイトと子供向けページ内の外部リンクの 2 つに注目する．

3.1. 子供向けポータルサイト

「Yahoo!きっず」や「キッズ goo」などの子供向けポータルサイトにはリンク集が存在する．「Yahoo!きっず」では約 63,000 件，「キッズ goo」では約 20,000 件のサイトが登録されており，これらのサイトを収集することで大量の子供向けサイトを収集できる可能性がある．しかし，紹介されているサイトすべてが子供向けではなく，一般向けの観光ガイドや公共機関などのサイトも存在する．したがって，サイトが子供向けであるかを判定し，テキストを収集する必要がある．

3.2. 子供向けページ内の外部リンク

子供向けページの外部リンクには，他の子供向けサイトをリンク先とするものが存在する．当研究室の事前調査[3]において，子供向けサイトを再帰的にクローリングすることで，大量の子供向けサイトが収集できる可能性を確認している．しかし，単調なクローリングではコーパスの構築に利用できない一般のサイトも大量に収集するため，ノイズを削減する必要がある．

これまで我々は，難易度推定システム「帯 2」[4]と子供向けサイトに含まれる特徴的なキーワードを用いて，サイトのトップページのみで子供向けサイトか判定し，ノイズの削減を試みた．しかし，トップページのみを利用した判定では，最も精度の高いキーワードによる判定手法でも 46.2%であり，半分以上のノイズが削減できていない．また再現率は 50.0%であり，抽出漏れも多い．

そこで本稿では，サイトのトップページだけでなく，そのサイトに含まれるページを単位とし，難易度およびキーワードの 2 つの手法と SVM（Support Vector Machine）による判定手法の 3 手法について比較・検討する．

4. 子供向けページ判定手法の検討

4.1. 難易度による判定手法

帯 2 は，各学年の教科書コーパスを活用することで，与えられたテキストの難易度を小学 1 年生から大学生以上までの 13 段階で推定する. 小島らの調査によると，小学生向けと明示されている Web ページの難易度の平均値は中学 1 年生であり，対象ページ 135 件中 65 件が中学 1 年生の難易度と推定されている．したがって，Web ページの難易度が中学 1 年生以下と推定された場合，子供向け Web ページとして収集できる可能性がある.

4.2. 特徴的なキーワードによる判定手法

子供向けサイトには，訪問者に子供向けであることを示すために，各ページに「こども向け」や「キッズページ」など子供に関連するキーワードが示されていることが多い. また，一般向け Web ページと子供向け Web ページを区別して管理するために，URL に「kids」などのキーワードを含んでいる場合もある. これらのキーワードがページ内や URL に含まれていれば, 子供向け Web ページとして収集できる可能性がある.

4.3. SVM による判定手法

子供向けサイトは，一般向けのサイトと比べて下記に示す特徴があると考えられる.

(1) 文章が平易で，理解しやすい.
(2) ふりがなが振られている.
(3) 漢字の量が少ない.
(4) 特徴的なキーワードが含まれている.
(5) 子供向け表現が多く，難解な表現が少ない.
(6) Flash，画像等を使用している.

(1)については，子供向けサイトは，訪問者として子供を対象にしているため，一般向けのサイトと比べて文章全体の難易度が平易であると考える．(2)と(3)については，子供が読めない漢字に対してふりがなを振る，難しい漢字は平易な語句に言い換えるなど，あまり利用されないと考えられる．(4)については，4.2 節で述べたように，訪問者に子供向けであることを示すため，本文や URL に特徴的なキーワードが含まれやすいと考えられる．(5)については，ページが子供に親しみやすいように文末に易しい表現を使用し，難解な表現が含まれる割合は少ない可能性がある．(6)については，子供に理解しやすいページを構成するために，一般向けサイトより Flash や画像等が多く使用されると考えられる.

(1)，(3)，(5)，(6)の特徴は，検索結果を子供向けにリランキングする岩田らの研究[5]でも注目しており，特に(5)はリランキングの精度の向上に寄与している．(1)～(6)の特徴を Web ページから抽出し，学習することで，子供向けサイトを判定できる可能性がある. そこで本研究では分類器に SVM を利用し，(1)～(6)を素性として組み合わせて子供向けページを判定する手法を提案する.

5. 評価手法

4 章で述べた 3 つの手法を用いて子供向けページを判定し，判定精度，再現率，F 値で比較する.

SVM の学習データは，Yahoo!きっず及び Yahoo!JAPAN に掲載されているサイトから人手で選択して構築する.

SVM による判定手法の性能は，*n* 分割交差検定法で評価する．難易度による判定手法と特徴的なキーワードによる判定手法については，*n* 分割したデータに対する平均値で評価する.

6. おわりに

本稿では，子供向けテキストを大量に収集した子供 Web コーパスを構築するために, 子供向けページを判定する手法について検討した.

今後は，これらの手法の性能を評価し，最も妥当である手法を選択する．そして，その手法を利用して，子供 Web コーパスを構築し，言い換え知識の抽出を行う.

謝辞

本研究の一部は，JSPS 科研費 25350335 の助成を受けて実施した.

参考文献

[1] NIE 実践の実態調査結果報告 http://nie.jp/inves/ji1_200807.pdf
[2] 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.
[3] 泉川 洸一郎，安藤一秋，“子供向け Web サイト収集のためのクローリング手法の検討”, 第 14 回情報科学技術フォーラム一般講演論文集，E-17, pp.231-232，2015.
[4] 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.
[5] 岩田 麻佑，荒瀬 由紀，原 隆浩，西尾 章治郎，“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.