Web一般新聞記事を子供向けに言い換える知識の抽出

藤沢 祐輔[*1], 相原 慎太郎[*2], 安藤 一秋[*3]
[*1] 香川大学大学院工学研究科 [*2, *3] 香川大学工学部
[*1] s10g483@stmail.eng.kagawa-u.ac.jp [*3] ando@eng.kagawa-u.ac.jp

1. はじめに

近年，小学校では，新聞を教材に用いる教育（NIE : Newspaper in Education）が実施されている[1]. しかし，新聞記事には小学生が読めない漢字や分かりにくい表現が存在するため，記事の内容を理解できない問題がある．新聞記事を小学生が理解できる表現に自動で言い換えることで，この問題が改善できると考えられる．そこで我々は，Web上の一般向け新聞記事（一般記事）を対象に，一般記事に含まれている小学生にとって難しい単語や表現を言い換える研究を進めている[2]. 難しい単語や表現を言い換えるための知識として，小学生向けの辞書が存在する．しかし，広辞苑に比べると語彙数は1/10程度しかなく，一般的な語しか掲載されていないため，記事を言い換える知識源としては網羅性に欠ける．

本研究では，Web上に存在する子供向け新聞記事（子供記事）に注目した．子供記事は，一般記事を分かりやすく書きなおしたものである．つまり，内容が一致する子供・一般記事のペアを収集し，差分を取ることで，子供向けに言い換えるための知識が抽出できると考えられる．これにより，小学国語辞典の語彙数の問題が改善できる．本稿では，まず，Web上の子供記事と一般記事を収集・分析することで，子供・一般記事のペアから言い換え知識が抽出できる可能性について調査する．そして，調査結果を基に，子供・一般記事のペアから言い換え知識を抽出する手法について検討する．

2. 事前調査

2.1 子供記事に対応する一般記事の割合調査

現在，毎日，朝日，読売新聞社などが，Web上で子供向けの新聞を公開している．そこで，公開されている子供記事に対して，内容が一致する一般記事がどの程度収集できるのかを調査する．対象とする子供記事は，最も更新頻度（公開記事数）が多い毎日小学生新聞とする．調査対象の記事は，2011年5月24日~29日の6日分54件である．なお，子供記事の内容が一般記事に半分以上出現した記事のみ対応付けることにする．調査は1名が人手で行う．

調査の結果, 子供記事は1日10件程度しか公開されていないことが確認できた．また，54件の子供記事の内，26件（約48%）は一般記事を対応付けることができた．内訳として，子供新聞1記事の内容が一般新聞1記事で対応付けできたものが15件, 子供新聞1記事の内容を対応付けるために複数の一般記事が必要であったものが11件であった．残りの28件（約52%）は, インタビューや天気図の解説など，子供記事独自の内容が書かれており，内容が一致する一般記事が存在しなかった．

2.2 対応記事間における文レベルの対応付け調査

対応記事から言い換え知識を抽出するためには，内容が一致する文同士を対応付ける必要がある．そこで，2.1で対応付けできた26件の子供記事に含まれる166文に対して，一般記事に含まれる文との対応付けを行い, 対応付けられる文の割合を調査する．なお，調査は1名が人手で行う．

調査の結果，子供記事に含まれる166文中112文（67%）が，それぞれ対応する一般記事の文に対応付けできた．その内，87文（52%）は1対1の関係で対応付けができた. 残りの文に関しては, 11文(7%)が1対2，6文（4%）は2対1，8文（5%）が2対2の関係で対応付けできた．全く対応付けできなかった54文（33%）は，新聞記事に現れる用語の解説，記者の感想など子供記事独自の内容であった．以上より，内容が一致する記事間においては，子供記事内の文の67%が何らかの形で一般記事内の文に対応付けできるため，言い変わっている箇所を調べることで，言い換え知識が抽出できる可能性がある．

2.3 子供・一般文ペア内の言い換え表現の調査

2.2で対応付けできた112文のペアに含まれる言い換え表現を調査する．

調査の結果，言い換えの単位として，単語単位とフレーズ単位の言い換えが存在することが確認できた．表1に言い換え例を示す．

表1. 人手で抽出した言い換え表現の例

子供文の表現	一般文の表現
単語単位（32件）	
被害	爪痕
見る	目撃する
日程	スケジュール
フレーズ単位（36件）	
修理するお金	修繕費
安全性に関する検査	安全性検査
体が分解されなかった	全身が残った

多くの言い換えでは，子供文の表現が一般文の表現に比べて簡単になっていたが,「日程」と「スケジュール」のように言い換えなくても理解できそうな単語や「分解されなかった」と「残った」のように，子供文の表現の方が難しいと思われる事例も数は少ないが確認された．

2.4 考察と今後の方針

まず，記事同士の対応付けにおいては，子供記事

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

には独自のテーマで書かれることがあるため，約半分の子供記事しか対応付けできないことがわかった．また，文レベルの対応付けにおいても，子供記事独自の内容が影響し，67%しか対応付けできないことがわかった．しかし，対応付けできた文からは，言い換え知識が抽出できることも確認できた．また，一部の例外は存在したが，対応付けした文から言い換え知識を抽出することで，やさしい語・表現への言い換え知識が構築できるといえる．

記事間で対応付けできなかった文でも，部分的な構造を見れば，Web 上に類似する文やフレーズが存在する可能性がある．これらを対応付けすれば，言い換え知識の抽出対象を広げることが可能である．新聞記事以外のテキスト利用は今後の課題として，まずは，対応する新聞記事に含まれる文のペアから言い換え知識を抽出する手法の実現を目指す．

また，言い換えの単位として，単語単位の言い換えに比べ，フレーズ単位の言い換えが若干多いことがわかった．まずは，抽出が容易であると考えられる単語単位の言い換え知識の抽出を目指す．なお，子供記事が1日あたり10件程度しか公開されないため，大量の記事を収集することが困難である．そこで，まずは表層情報を利用した規則ベースの手法について検討する．

3. 単語単位の言い換え知識の抽出

2. の事前調査を基に，子供・一般記事のペアから単語単位の言い換え知識を抽出する手法を以下に定義する．

【手順 1】子供・一般新聞社サイトから記事を収集
【手順 2】記事間の類似度を計算し，子供記事と内容が一致する一般記事の対応付け
【手順 3】文間の類似度を計算し，子供・一般記事ペアに含まれる文の対応付け
【手順 4】文のペアから言い換え表現対を抽出
【手順 5】抽出した言い換え表現対の妥当性を検証

3.1 子供・一般記事群の収集[2]

言い換え知識を抽出するために，同じ内容が記述された一般，子供記事を収集する必要がある．Web 上で公開されている子供新聞は，複数の新聞社を合わせても 1 日 10 件程度しかないため，子供記事を収集後，それに対応する一般記事を収集する方が効率がよい．本研究では，毎日と朝日，読売小学生新聞の各サイトから子供記事を自動収集して利用する．

収集法について説明する．収集した子供記事から重要語を抽出し，それを基に Web 検索する．そして，検索結果に含まれる子供記事を除いた上位 10 件を子供記事に対応付ける一般記事候補群とする．タイトルと 1 文目の自立語の中から，tf・idf 値を計算して重要語を抽出する．

3.2 記事同士の対応付け[2]

一般記事候補群から子供記事の内容に最も近い記事同士を対応付ける．具体的には，両記事のタイトルと記事本文に含まれる自立語を基に Jaccard 係数を利用して類似度計算し，類似度が最も高い記事同士を対応付ける．なお，複数の記事への対応付けは今後の課題とする．

3.3 文同士の対応付け[2]

対応付けた記事のペアから文の内容が最も近い文同士を対応付ける．具体的には，文字 3-gram の一致数が最も高い文同士を対応付ける．

3.4 言い換え表現対の抽出

対応付けた文のペアから自立語の言い換え候補の抽出を行う．中村ら[3]は，対応付けられた文のペアから同義語，広義語や語の位置を手掛かりとして，文中の語に対する言い換え対を抽出する手法を提案した．また，山崎ら[4]は，係り元と係り先の文節を制約として利用し，共通の係り元と係り先に挟まれる文節の名詞句を換言知識として抽出する手法を提案した．しかし，いずれの手法とも名詞（句）のみを対象にしていた．本研究では，文節の修飾・被修飾の関係を基に言い換え個所を絞り込み，自立語の言い換え候補を抽出する手法を検討する．

本研究で利用する子供・一般記事に含まれる文のペアは意味的類似度が高いため，Web 上の一般テキストを対象とした山崎らの手法より抽出の制約を緩くできる．したがって，対応付けした文のペアにおいて，共通して出現する自立語（共通自立語）を含む文節が修飾している語または共通自立語を含む文節に修飾される語を言い換え候補として抽出する．

子供記事の文（子供文）は，一般記事の文（一般文）をわかりやすく表現するために，文節の追加や削除が行われる．そこで，子供文と一般文の修飾・被修飾の関係を効率よく比較しながら言い替え表現候補を抽出するために，係り受け関係から部分文を生成して利用する．例えば，「日本代表が/カメルーンに/勝利しました.」からは，「日本代表が/勝利しました.」と「カメルーンに/勝利しました」の部分文が生成される．

以下，部分文を利用した言い換え候補の抽出手法の概要を示す．

① 子供・一般文ペアから共通自立語を抽出する．なお，文内の自立語は複合語化して利用する．
② 各文に対して係り受け関係を求める．
③ 各文の修飾・被修飾関係を基に部分文を生成する．
④ 共通自立語が含まれる部分文同士を対応付ける．
⑤ 共通自立語を含む文節に対して，共通自立語を含む文節を修飾している自立語または共通自立語含む文節に修飾されている自立語を一般化して抽出する．

次に，子供文と一般文から言い換え候補を抽出する例を示す．

（子供文）オリンピックの開催に向け A 建設会社が新しい橋を上海に作った
（一般文）五輪の開催に向け A 建設会社によって上海の中心部に新しい橋が建設された
① 子供・一般文ペアから共通自立語を抽出する．
共通自立語：開催，向け，A 建設会社，新しい，

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

橋，上海
②，③ 係り受けを基に部分文を生成する．
子供文の部分文の例：
　オリンピック・の→開催・に→向け，→作っ・た
　A・建設・会社・が→作っ・た
　新しい→橋・を→作っ・た
　上海・に→作っ・た
一般文の部分文の例：
　五輪・の→開催・に→向け→建設・さ・れ・た
　A・建設・会社・によって→建設・さ・れ・た
　上海・の→中心部・に→建設・さ・れ・た
　新しい→橋・が→建設・さ・れ・た
④ 共通自立語が含まれる部分文同士を対応付ける．
＜開催・向け＞
（子供文）オリンピック・の→開催・に→向け
　　→作っ・た
（一般文）五輪・の→開催・に→向け
　　→建設・さ・れ・た
＜新しい・橋＞
（子供文）新しい→橋・を→作っ・た
（一般文）新しい→橋・が→建設・さ・れ・た
⑤ 共通自立語を修飾している表現を一般化して抽出する．
（子供文）オリンピックの→開催・に
（一般文）五輪・の→開催・に
言い換え候補：五輪→オリンピック
　または，共通自立語に修飾されている表現を一般化して抽出する．
（子供文）橋・を→作っ・た
（一般文）橋・が→建設・さ・れ・た
言い換え候補：建設する→作る

3.5　妥当性検証

3.4 で抽出した言い換え候補ペアに対して検索エンジンのヒット数を用いて妥当性を検証する．

子供文と一般文の難易度は，子供文 ＜ 一般文であると考えられる．そこで，本研究では難易度の低い語は汎用性が高い(ヒット数が多い)と仮定し，以下の 2 つの手法を比較検討する．

【手法 1】子供文から抽出された言い換え候補（子供語候補）のヒット数と一般文から抽出された言い換え候補（一般語候補）のヒット数が，子供語候補のヒット数 ＞ 一般語候補のヒット数という関係の場合，そのペアを言い換え関係にあるとして抽出する．

【手法 2】一般語の方が汎用的な場合も考えられるため，手法 1 にヒット数が激減しないという条件を加え，いずれかが満たされた場合，言い換え関係にある語として抽出する．

以下，検索エンジンのヒット数を利用した検証手順を示す．

① 一般語をクエリとして検索し，検索結果の上位 *n* 件の Web ページを取得する．

② *n* 件の Web ページから，2 タイプの文節 2-gram，係り元の文節+一般語候補を含む文節（「前文節+一般語」とよぶ）と一般語候補を含む文節+係り先の文節（「一般語+後文節」とよぶ）の出現頻度をカウントし，上位 *m* 件ずつの文節 2-gram を抽出する．なお，以下に，2 タイプの文節 2-gram の例を示す．
　前文節+一般語：熱波の　米国
　一般語+後文節：米国は　主張

③ 各 *m* 件の文節 2-gram をクエリとして検索し，それぞれのヒット数 *hit* を得る．
hit(前文節+一般語) = *hit*(“熱波の米国”) = 1,000
hit(一般語+後文節) = *hit*(“米国は主張”) = 2,900

④ ②で求めた各 *m* 件の文節 2-gram に対し，一般語を子供語に言い換えて，③と同様に *hit* を求める．
hit(前文節+子供語) = *hit*(“熱波のアメリカ”)= 2,200
hit(子供語+後文節) = *hit*(“アメリカは主張”)= 3,800

⑤ ③と④で求めた *hit* 値を用いて，妥当性検証のためのスコア *score* を求める．

$$score(before) = \frac{hit(\text{前文節} + \text{子供語})}{hit(\text{前文節} + \text{一般語})} = 2.2$$

$$score(after) = \frac{hit(\text{子供語} + \text{後文節})}{hit(\text{一般語} + \text{後文節})} = 1.3$$

⑥ ⑤で求めた *score* を利用して，2 つの手法で言い換え表現対 p_pair_i の妥当性検証を行う．

【手法 1】
　合わせて 2*m* 件の $score(before_i)$と $score(after_i)$の平均が 0 より高ければ，妥当な言い換え表現対と判定する．

$$validity(p\_pair_i) = \frac{\sum score(before_i) \sum score(after_i)}{2m}$$

【手法 2】
　全体で 2*m* 件の *score* が αを超える個数を数え，その個数が βを超えた場合，妥当な言い換え表現のペアと判定する．なお，予備実験により，α=0.1，β= 5 として利用する．

4.　評価実験

4.1　新聞記事の収集に対する評価

数が限定される子供記事は網羅的に収集可能であるため，一般記事の収集について評価する．無作為に抽出した 20 件の子供記事に対して，Web 検索で得られる各上位 10 件の一般記事の内容を人手で判断し，子供記事の内容に一致する記事の平均含有率を調査する．

評価の結果，平均含有率は 50%であった．対応記事が収集できなかった子供記事を分析した所，一般記事には存在しない子供記事独自の話題を扱ったもの（4 件）や複数の一般記事の内容で構成されたもの（2 件）が存在した．前者は対応付けできないため，フィルタリング手法を考案する必要がある．後者は，複数存在する可能性があるため，取り扱い方を検討する必要がある．これらの記事を除いた子供記事（14 件）に対する平均含有率は 71%であった．今後は，タイトルだけでなく，本文情報を利用する方法も検討する．

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

4.2　記事単位の対応付けに対する評価

10 件の子供記事に対して，Web 検索で得られる各上位 10 件の一般記事の内容を人手で判断し，子供記事の内容に一致する一般記事が現れる最高順位の平均値と，子供記事の内容に関連する一般記事が現れる最高順位の平均値で評価する．評価データには複数の一般記事から構成されたと考えられる子供記事は利用していない．

評価の結果，内容が一致する記事同士の平均順位は 1.3 位，関連する記事同士の平均順位は 1.0 位となり，Web 検索で子供記事に対応する一般記事が収集可能であることが確認できた．

4.3　文単位の対応付けに対する評価

子供・一般記事のペア 10 件に対し，内容が一致する子供・一般記事の文同士を人手で対応付けした結果，子供記事の総 80 文に対し，58 文が対応付けできた．これを正解文として，提案手法で対応付けた文と比較し，再現率を求めることで評価する．

評価の結果，再現率は 89.7%であった．対応付けができなかった原因を分析すると，文字 3-gram を利用したため，一般文と子供文で語順が入れ替った文や別の表現に言い換えられた文などが存在した．

4.4　抽出した言い換え表現対の妥当性検証の評価

提案手法で抽出した言い換え候補に対して妥当性検証を行い，最終的に抽出された言い換え表現対の精度を求めて評価する．評価に利用する記事は，よみうり博士のアイデアノート[5]から抽出した 96 件の一般・子供新聞記事ペアである．また，山崎らの手法[4]を用いて言い換え表現を抽出し，本手法の妥当性検証を適用した場合の精度とも比較する．

評価結果を表 2 に示す．提案手法および比較手法において，手法 1，2 を比較すると，2 つの閾値を用いた手法 2 の精度が共に高くなった．これにより，言い換え後のヒット数が激減しないという制約が有効に働いたといえる．また，提案手法は山崎らの手法と比べて，抽出のための制約が緩いため精度が低くなった．しかし，抽出できた言い換え表現対を比較すると，提案手法が約 4 倍も多い．

表 2．評価結果

表現対の抽出手法	提案手法		比較手法	
妥当性検証の手法	手法 1	手法 2	手法 1	手法 2
抽出した表現対数	91	80	19	15
正しい表現対	57	53	16	13
精度	0.63	0.66	0.84	0.87

提案手法で抽出できなかった言い換え表現対を分析した結果，以下の特徴が確認できた．

① 言い換えの範囲がフレーズ

子供文：白菜の→値段が→高くなった.

一般文：白菜が→高騰した.

② 子供文では難しい漢字がひらがな（読み仮名）で表現される場合があり，係り受け解析が失敗

子供文:団地にあるしき地約 15 ヘクタールに工場を建設

係り受け解析の結果（誤り）

団地に→ある→しき→建設

地約 15 ヘクタールに→建設

③ 助詞を考慮しないことが影響

子供文:愛媛県教育委員会が 2010 年，新設した「えひめ授業の鉄人」に 5 人の先生が選ばれました.

一般文：県教委が今年度，新設した「えひめ授業の鉄人」に 5 人の先生が選ばれた.

子供文の部分文：先生が選ばれました.

一般文の部分文：今年度，選ばれた.

「先生」と「今年度」が言い換え表現対として抽出されてしまう.

問題①は，言い換えの抽出対象をフレーズに拡張することで対応できる．フレーズ単位の言い換えに対応できた場合，抽出できる言い換え表現対は 18 個増え，精度は 0.89 まで向上できる．今後は，フレーズに対応した言い換え知識の抽出手法を考案する.

また，問題②は，読み仮名を漢字へ変換することで解析ミスを防ぐことができる．問題③については，助詞を考慮した対策を考える必要がある.

5.　おわりに

本稿では，Web 上の子供・一般記事を収集・分析することで，子供・一般記事のペアから言い換え知識が抽出できる可能性について調査した．調査の結果，単語・フレーズ単位の言い換え知識が抽出できる可能性を確認した．そして，調査結果を基に，係り受け関係を基に単語単位の言い換え知識を抽出する手法を検討した．抽出した言い換え知識の妥当性を評価した結果，精度は 66%となった．今後は，まず，フレーズ単位の言い換え知識を抽出する手法を考案する予定である.

謝辞

本研究は，文部科学省科学研究費補助金（若手研究(B) 22700813）と平成 22 年度香川大学若手研究経費の助成を受けて実施した.

参考文献

[1]教育に新聞を，http://www.nie.jp/

[2]藤沢，安藤：“係り受け関係を利用した一般新聞記事を子供向けに言い換える知識の抽出”，第 9 回情報科学学術フォーラム，pp.299-300 (2010)

[3]中村，田中，北野，田中，大林，“児童向け新聞教材のための言い換え表現対の抽出に関する研究”，情報処理学会第 71 回全国大会，3S-1，pp.2-293-2-294　(2009)

[4]山崎，沢井，山本，“構文情報を用いた名詞句の換言”，言語処理学会第 12 回年次大会，B4-6，pp779-782，(2006)

[5]よみうり博士のアイデアノート，http://www.yomiuri.co.jp/nie/note/top.html

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved