E-035

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

Extraction of Knowledge for Paraphrasing Newspaper Articles for Children using Dependency Relations

藤沢 祐輔† 安藤 一秋‡
Yusuke Fujisawa Kazuaki Ando

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[1,2]．しかし，新聞には子供が読めない漢字や分かりにくい表現があるため，子供が新聞の内容を理解できない問題がある．そこで，Web 上の一般向け新聞記事を子供向けの表現に自動で言い換えることができれば，この問題を改善できる．

そこで本稿では，一般新聞記事を子供向けに言い換える知識の抽出法を検討する．具体的には，Web 上に存在する子供向け新聞記事（子供記事）とその内容に一致する一般向け記事（一般記事）を収集し，文レベルで対応付ける．そして，対応付けられた文のペアから係り受け関係を利用して，自立語言い換え候補を自動抽出する手法を検討する．

2. 言い換え候補の抽出

2.1. 子供新聞記事の収集

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．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．したがって，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．本稿では，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する．

2.2. 一般新聞記事の収集

一般記事は検索エンジンで収集する．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．

2.3. 記事単位の対応付け

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．両記事のタイトルと記事本文に含まれる自立語を基に，ベクトル空間モデルで対応付ける．類似度計算には，Jaccard 係数を利用し，類似度が最も高い記事同士を対応付ける．

2.4. 文単位の対応付け

対応付けし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．

2.5. 言い換え候補の抽出

対応付けた文のペアから動詞，名詞，形容詞の言い換え候補の抽出を行う．中村ら[2]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に着目して名詞句を換言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．本稿では，山崎らの手法と同様に，係り受け解析によって得られる文節の修飾・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．

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．部分文を生成することで，文節の出現順による影響も吸収できる．

また，山崎ら[4]の手法と同様，文のペアにおいて，共通自立語が修飾している語，共通自立語に修飾される語は，同義関係の可能性があると仮定し，部分文に現れる共通自立語と係り受け関係にある語を言い換え候補として抽出する．以下，言い換え候補の抽出手法の概要を示す．

① 子供・一般文ペアの両方に含まれる自立語（共通自立語）を抽出する．
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 共通自立語に対して，共通自立語を修飾している自立語または共通自立語に修飾されている自立語を抽出する．

＜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。

⑥ 検索エンジンのヒット数を用いて，言い換え候補の妥当性を検証する．

次に，子供記事の文（子供文）と一般記事の文（一般文）から言い換え候補を抽出する例を示す．

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，橋，上海

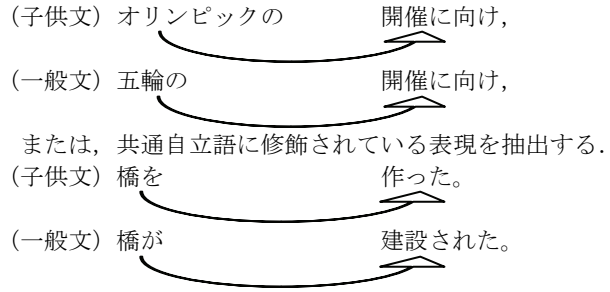
②，③ 係り受けを基に部分文を生成する．
子供文の部分文の例：
オリンピック・の→開催・に→向け・，→作っ・た・。
A・建設・会社・が→作っ・た・。
新しい→橋・を→作っ・た・。
上海・に→作っ・た・。
一般文の部分文の例：
五輪・の→開催・に→向け・，→建設・さ・れ・た・。

† 香川大学大学院工学研究科，‡香川大学工学部

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抽出する．



または，共通自立語に修飾されている表現を抽出する．

⑥ 検索エンジンのヒット数を用いて，言い換え候補（五輪→オリンピック, 建設する→作る）の妥当性を検証する．

3. 対応付け手法の評価

3.1. 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するため, 一般記事の収集精度について評価する．無作為に抽出した 90 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, 子供記事の内容に類似する記事の平均含有率を調査する．

実験の結果，50 件の子供記事に対して内容が類似する一般記事が収集でき，平均含有率は，55.6%となった．一般記事が収集できなかった子供記事を分析した所, 一般記事に存在しない子供記事独自の話題で多く存在した. 独自の話題を扱った子供記事については, 記事単位の対応付け時に類似度が低くなるため, 機械的に排除可能である．また，数は多くないが，検索結果の上位 10 件に内容が類似する一般記事が存在しない場合もあった. 記事単位の対応付けでノイズとなる記事を排除できるため, 収集する一般記事候補群の数を増やすことで対応できる．

3.2. 記事単位の対応付けに対する評価

記事単位の対応付け手法の有効性を評価する．まず，子供記事 50 件に対して, それぞれ上位 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 1 件の一般記事をそれぞれ正解データとする．そして，本手法で記事の対応付けを行い，スコア順に整列して，正解データが現れる順位の平均順位を求める．

実験の結果，平均順位は 1.78 位であった．TF・IDF を利用して類似度を計算しているため, 文章が少なすぎる子供記事は, 内容が一致する一般記事より関連する内容の記事が上位に出現してしまうことがあった．

3.3. 文単位の対応付けに対する評価

文単位の対応付け手法の有効性を評価する．子供・一般記事のペア 50 件に対し，文のペアを抽出する．そして，抽出した子供文 471 文に対し, 同じ内容の文を対応付けているかを人手で判定し，精度を求める．

実験の結果, 136 文が正しく対応付けられ, 精度は 28.9%であった．精度が低い原因は，内容が一致する文だけでなく，類似した文からも言い換え知識を抽出するために，フィルタリングしていないことが影響している．また，子供記事の文と一般記事の文が完全に対応しない場合もある．さらに，子供向けにわかりやすく説明するために，補足として追加された文やほとんどの文節が言い換えられた子供文などがあり，その対応付けに失敗した．

3.4. 言い換え候補の抽出に対する評価

3.3 で対応付けた 136 文のペアに対して，人手で言い換え表現対を抽出し，これを正解データとする．そして，本手法で抽出した言い換え候補対と正解データを基に再現率を求める．但し，ここで利用する言い換え候補対は，候補抽出のステップ⑥が検討中であるため, ステップ⑤の出力を利用する．なお，正解データは 23 個であった．

実験の結果，本手法で抽出した言い換え候補対は 15 個で，再現率は 65.2%となった．抽出できなかった言い換え表現を分析すると，共通自立語が言い換えられていたり，「～すること」や「～するつもり」などのことが名詞として扱われており, 共通自立語として正しいものが取れない場合などが確認できた．また，フレーズ単位で言い換えられている表現もあり, そこに含まれる言い換え表現が抽出できなかった．最後に，抽出された表現対の例を示す．

＜子供向け＞		＜一般向け＞
アピールする	→	力説する
示す	→	示唆する
オリンピック	→	五輪
減る	→	減少する
持つ	→	保有する
落ちる	→	落下する
話した	→	語った
下がる	→	下落する

4. おわりに

本稿では，子供・一般向け記事の対応付け手法と文単位の対応付け，自立語言い換え候補の抽出手法を検討した．評価の結果，共通自立語と係り受け関係を用いて，名詞・動詞の言い換え表現を抽出することができた．今後は，共通自立語が言い換えられているパターンに対応する手法を考案する．また，検索エンジンのヒット数を用いて，言い換え候補の妥当性を検証する方法を考案する. 最終的には，言い換え辞書と文脈情報を基に，一般向けのニュースから子供向けの表現に自動で言い換える手法を考案する．

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] 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).
[3] 乾，藤田，“言い換え技術に関する研究動向”，自然言語処理，Vol. 11, No. 5, pp. 151-198 ，(2004)
[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)