4I1-R-9-7

小学生を対象としたWeb新聞読解支援のための説明語抽出の検討

Keyword Extraction from Web Newspapers for Elementary School Students

小林健[*1]
Ken Kobayashi

安藤一秋[*2]
Kazuaki Ando

[*1] 香川大学大学院工学研究科
Graduate school of Engineering, Kagawa University

[*2] 香川大学工学部
Faculty of Engineering, Kagawa University

Recently, NIE (Newspaper In Education) has become increasingly active at elementary schools. Elementary school students can improve the ability of communication and reading by NIE. However, there are Kanji characters or terms which students cannot read and understand in newspapers. In order to increase the effectiveness of NIE, educational institutions require a support system for reading newspapers. This paper discusses a method for extracting keywords which should be explained for elementary school students from Web newspapers.

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[NIE 12]）が実施されている．しかし，小学生は語彙が少ないため，新聞を読むことが難しい．また，文章の読解力にも個人差があるため，文章を理解できたとしても，その理解度に差が生じる．したがって，NIE の効率的な実施を妨げてしまう可能性がある．そこで，新聞の読解を支援できれば，NIE の効果をより高めることができると考えられる．

小学生の新聞読解を支援するツールの一例として，読売新聞が Web 上で公開している「よみうり博士のアイデアノート[読売 12]」がある．機能として，用語説明や関連情報の提示などが実装されているが，説明の付与や提示する情報の整備は人手で行われている．そのため，更新頻度は 1ヶ月に 1 記事程度で，利用できる記事数がかなり少ない．

そこで本研究では，NIE の効率を高めるために，Web 新聞記事に対して読解支援を行うツールの構築を目的とする．本稿では，その第一段階として，小学生に説明すべき重要語を抽出する手法について検討する．

2. NIE 実践校へのインタビュー

読解支援の方針を検討するため，平成 22 年度まで NIE 実践校の指定を受けていた高松市立屋島小学校の NIE 担当教員にインタビューを実施した．

その結果，以下の点が明らかになった．

(1) 新聞は難しい表現や単語が多く分かりにくい．
(2) 多くの児童が新聞を探すことすら難しい．
(3) 教師の負担（記事選択等）が大きい．
(4) 新聞記事に関連する写真やグラフがほしい．

(1)を解決するため，当研究室では新聞記事中の難しい表現をわかり易い表現に言い換える研究[藤沢 12]に取り組んでいる．また，(2)と(3)を解決するために，記事推薦に関する研究[坪井 11]を進めている．本研究では，(1)と(4)を解決するため，新聞の読解を支援するためのシステム構築を行う．

連絡先：小林健，香川大学大学院工学研究科，香川県高松市林町 2217−20，s11g474@stmail.eng.kagawa-u.ac.jp

3. 読解支援システムの概要

現在検討している読解支援システムの機能を以下に示す．

(a) 重要語に対する自動説明付与
(b) 記事に関連する数値データや写真の提示
(c) 複数の新聞記事の比較

(a)〜(d)の機能概要を説明する．(a)は，記事中の重要語に対して，簡単な説明を自動で付与する機能である．(b)は，小学生が新聞をより深く読み進めるために必要なグラフや図などを Web 資源から探し出して提供する機能である．(c)は，同じ内容を扱う異なる新聞社の記事を提示することで，読み比べを支援する機能である．また，関連記事の提示も行う．

本稿では，これらの機能のうち(a)を実装するための第一段階として，重要語の抽出手法について検討を行う．

4. 重要語の抽出手法

小学生は語彙が少ないため，新聞を読むことが難しい．よみうり博士のアイデアノートでは，記事の主題に関係する重要語の他に，専門用語や難しい語に対し，手動で説明を付与している．本研究においても，よみうり博士のアイデアノートの考え方を基準として，主題に関係する重要語と専門用語や難しい語などに対して，説明を付与する．

本稿では，まず，後者に焦点をあて，難易性と専門性の両方をもつ単語を説明すべき重要語（説明語）として定義する．また，難易性の指標として難易度を，そして，専門性の指標として専門度を定義し，説明語を抽出する．

4.1 難易度

難易度とは，記事中に出現する単語の難易性を推定した指標として定義する．本研究では，難易度を親密度と抽象度を用いて推定する．

我々大人でも普段あまり見ない，馴染みのない単語は，わかりにくい場合が多い．この点は小学生にとっても同じであるといえる．そこで，難易度計算の 1 つ目の指標として単語の馴染みの程度を表す「親密度」を利用する．

また，小学生は語彙が少ないため，抽象的な単語より，意味が細かく，具体的な単語の方が習得していない場合が多いと考えられる．そこで，2 つ目の指標として，単語の意味の広さを表す「抽象度」を利用する．以下，親密度と抽象度について説明する．

(1) 親密度

親密度は日本語の語彙特性[天野 08]に掲載されている指標の 1 つである，「単語親密度」を用いて算出する．単語親密度とは，単語の馴染みの程度を数値で表した指標で，7 つの値（1：馴染なし～7：馴染みあり）で定義される．

例えば，「市場」と「逕庭」対する単語親密度は，「市場」= 6.188,「逕庭」= 1.5 となる．単語親密度が低い単語の方が難しい単語であるといえる．そこで，本研究では，単語 w の親密度 $sim(w)$を次式で定義する．

$$sim(w) = 1 - termsim(w)/7 \quad (1)$$
$$termsim(w) \ : \ w\text{の単語親密度}$$

ここで，7 は単語親密度の最大値である．

(2) 抽象度

抽象度とは，単語の意味がもつ広さを数値で表したものとして定義する．抽象度を算出するために，本研究ではシソーラスの深さを用いる．シソーラスとは，単語の上位/下位関係を階層構造で表現した辞書である．そのため，下位階層の単語ほど具体的な意味を持つ．

例えば，日本語語彙体系[池原 99]において，「経験」と「猛省」に対するシソーラス上の深さは，「経験」= 4,「猛省」= 9 となる．深さが深い「猛省」の方がより具体性を持つため，難しい単語である．したがって，シソーラス上の深さが深い単語を高難易度とする．

単語 w の抽象度を次式で定義する．

$$abs(w) = depth(w)/13 \quad (2)$$
$$depth(w) : \text{シソーラス上の深さ}$$

ここで，13 は日本語語彙体系の深さの最大値である．

(3) 難易度の算出

前述した親密度と抽象度を用いて，記事 D に含まれる単語 w_i の難易度 $dif(w_i)$を次式で定義する．

$$dif(w_i) = (1-\alpha_{dif})\frac{sim(w_i)}{sim\max(D)} + \alpha_{dif}\frac{abs(w_i)}{abs\max(D)} \quad (3)$$

ここで，

$simmax(D)$：D 内の $sim(w_i)$の最大値，

$absmax(D)$：D 内の $abs(w_i)$の最大値，

$\alpha_{dif}(0 \leqq \alpha_{dif} \leqq 1)$：重み

である．

4.2 専門度

専門度は，単語の専門性を表したものとして定義する．本研究では，専門度を計算するために，FLR[中川 03]と MDP[久保 10]の 2 つの手法について検討する．

(1) *FLR*

FLR[6]は，単語の連接頻度を利用する手法である．これは，多くの複合名詞を構成している用語は，専門用語であるという仮説に基づいている．そのため，多くの複合名詞に共通して存在する単名詞のスコアが高くなる傾向がある．

以下に，FLR の計算式を示す．

$$FLR(W) = f(W) \times \prod_{i=1}^{n} (FL(w_i)+1)(FR(w_i)+1)^{\frac{1}{2n}} \quad (4)$$

ここで，

w_i：W を構成する単名詞，

n：W を構成する単名詞の数，

$FL(w_i)$：w_i の左に名詞が連接する頻度，

$FR(w_i)$：w_i の右に名詞が連接する頻度，

である．

FLR を用いた専門度のスコア $spe_{FLR}(W)$は，$FLR(W)$を正規化して利用する．以下に式を示す．

$$spe_{FLR}(W) = \frac{FLR(W)}{FLR\max(D)} \quad (5)$$

ここで，$FLRmax(D)$は，記事 D 内の FLR(W)の最大値である．

(2) *MDP*(Minimum of the Difference between Population Proportions)

MDP[7]は，複数の分野コーパスを利用して，対象分野と対象分野以外のコーパスにおける単語の出現率の差を計算することで，単語が対象分野の専門用語であるかどうかを判断する手法である．

MDP は以下で計算される．

$$MDP(W) = \min Z_i \quad 1 \leq i \leq N \quad (6)$$

$$Z_i = \frac{\frac{f_0(W)}{W_0} - \frac{f_i(W)}{W_i}}{\sqrt{\pi_i(W)(1-\pi_i(W))(\frac{1}{W_0}+\frac{1}{W_i})}} \quad (7)$$

ここで

W：単語

N：他分野コーパスを構成する分野の数

W_0：対象分野に出現する単語の総延べ語数

W_i：i番目の他分野コーパスの総延べ語数

f_0：単語Wの対象コーパスでの出現頻度

f_i：i番目の他分野コーパスにおける単語Wの出現頻度

$$\pi_i(W) = \frac{f_0(W)+f_i(W)}{W_0+W_i} \quad (8)$$

MDP の値が高い程，対象分野の専門用語である可能性が高い．ここで，*MDP* は対象分野における単語 T の専門性を計算する手法であるため，対象となる分野が N 個あった場合，単語 T の *MDP* は N 個計算される．そこで本研究では，N 個計算された *MDP* の内，最も高い値を $MDP(W)$として利用する．

$MDP(W)$を用いた専門度のスコア $spe_{MDP}(W)$は，*FLR* の場合と同様，$MDP(W)$を正規化して利用する．

以下に，計算式を示す．

$$speMDP(W) = \frac{MDP(W)}{MDPmax(D)} \quad (9)$$

ここで，$MDPmax(D)$は，記事 D のすべての単語における *MDP* の最大値である．

(3) *MDP* の計算に利用するコーパス

MDP の計算には異なる分野のコーパスを利用する．Web 新聞記事は，各新聞社が独自の分野で管理している．そこで本研究では，小学生向けの分野を設定して利用する．分野の設定には，よみうり博士のアイデアノートと読売新聞社のカテゴリを参

考に,「社会」,「スポーツ」,「政治」,「歴史」,「教育」,「国際」,「科学」,「環境」,「経済」,「情報」の 10 分野とする. そして, Web 新聞記事を新聞社サイトから自動収集し, それらを 10 分野に分類することで, 分野コーパスを生成する.

本稿では, Web 新聞記事として, 読売新聞社のホームページから 2011 年 5 月に収集した 2891 件の新聞記事を利用する. なお, 記事の分類には, コンプリメントナイーブベイズ[Rennie 03]を利用する.

4.3 説明語の抽出

前述の専門度と難易度を用いて, 重要度を抽出するためのスコア *score* を計算し, 値が上位の語を説明語として抽出する.

記事 D に含まれる単語 w_i のスコア $score(w_i)$ を次式で定義する.

$$score(w_i) = (1-\alpha_{score}) \times dif(w_i) + \alpha_{score} \times spe(w_i) \qquad (10)$$

ここで, α_{score} は重み ($0 \leqq \alpha_{score} \leqq 1$) である.

5. 評価

5.1 評価方法

よみうり博士のアイデアノートに掲載されている記事を用いて提案手法を評価する. 評価には 2 つのデータセットを利用する.

データセット 1:6 年生の記事 60 件

データセット 2:6 年生の社会カテゴリに属する記事 46 件

正解データには, よみうり博士のアイデアノートで定義されている赤と青の強調単語を利用する. 青で強調された単語 (青キーワード) は, 単語の説明が付与されているだけでなく, 関連情報や教科書へのひも付けなど, 新聞をより深く読むための補足情報が付与されている. したがって, 本研究では, 青キーワードを優先的に抽出すべき単語と仮定する. なお, 各記事において正解データ数は異なる.

本研究では, 青キーワードの正解率と, 正解データ全体における正解率をそれぞれ計算し, 評価する. ここで, 正解率は, 式(11)により求める.

$$\text{正解率} = \frac{\text{抽出された正解の個数}}{\text{正解単語の個数}} \qquad (11)$$

スコア計算式(10)の係数 α_{score} を変化させながら, スコアの上位 M 件の単語を抽出し, 正解データと比較することで, 正解率を求める. この時, 専門度として, *MDP* と *FLR* をそれぞれ利用した手法の正解率を比較する. 今回の評価では, 新聞記事単位で, *FLR* を計算する. また, *MDP* は, 提案しているカテゴリを利用する手法と, 読売新聞のカテゴリをそのまま利用した手法の 2 つについて比較する. ここで, M は, 正解データ数である. なお, (10)式の係数 α_{score} は予備実験により 0.3 とする.

5.2 評価結果と考察

(1) データセット 1 による評価

データセット 1 に対する, 評価結果を表 1, 表 2 に示す. それぞれ, 表 1 が, 正解データ全体に対する正解率, 表 2 が青キーワードに対する正解率を示している.

表 1 より, α_{score} =0.2 の時, 読売新聞のカテゴリを利用した *MDP* の正解率が 0.537 で最大となった. また, 提案したカテゴリを採用した場合の *MDP* の正解率は α_{score} が 0.1 の時, 0.533 となった. これらを比較すると, 今回の評価では, 読売新聞のカテゴリをそのまま利用した方が, 提案したカテゴリを利用するよりも若干良い値を示している. しかし, その差は僅かである. さらに, 表 3 より, 青キーワードの正解率に関しては, 提案手法が α_{score} =0.2, 0.4 の時, 最大値 0.539 となっているのに対して, 読売新聞のカテゴリをそのまま採用した場合は最大値でも, α_{score} =0.3 の時 0.513 となり, 提案カテゴリの正解率が約 2 ポイント高い結果を示している. 青キーワードは, 調べ学習の起点となる重要な単語であることを考えると, 提案カテゴリを利用した方が, 読売新聞のカテゴリをそのまま利用するよりも, より, 重要な単語を抽出できる可能性が高いと考えられる. 今回, 提案したカテゴリは読売新聞のカテゴリとよみうり博士のアイデアノートを参考にしている. そのため, 結果的に, 似たようなカテゴリになってしまったと考えられる. そこで, 更に, カテゴリに工夫を加えることで, 正解率の向上が期待できる.

表 1: 全キーワードの正解率

α_{score}	*MDP* (読売新聞)	*MDP* (提案カテゴリ)	*FLR*
0	0.516	0.516	0.516
0.1	0.531	0.533	0.519
0.2	0.537	0.531	0.508
0.3	0.522	0.526	0.508
0.4	0.520	0.517	0.503
0.5	0.516	0.522	0.503
0.6	0.513	0.520	0.478
0.7	0.509	0.519	0.462
0.8	0.505	0.513	0.441
0.9	0.494	0.502	0.405
1	0.455	0.466	0.369

表 2:青キーワードの正解率

α_{score}	*MDP* (読売新聞)	*MDP* (提案カテゴリ)	*FLR*
0	0.480	0.480	0.480
0.1	0.487	0.500	0.493
0.2	0.500	0.539	0.513
0.3	0.513	0.533	0.533
0.4	0.493	0.539	0.539
0.5	0.500	0.520	0.539
0.6	0.500	0.513	0.507
0.7	0.493	0.520	0.507
0.8	0.480	0.513	0.526
0.9	0.474	0.513	0.493
1	0.454	0.480	0.434

続いて, *MDP* と *FLR* の比較を行う. α_{score}=0.1 の時, *FLR* を利用した手法の正解率が 0.519 となり, *MDP* の方が約 1~2 ポイント良い結果を示した. これは, 今回の評価では, 記事単位で *FLR* を計算したため, *FLR* よりも *MDP* の方が特定の分野に突出して表れるという専門用語の特性を良く表すことができたためだと考えられる. しかし, *FLR* と *MDP* のどちらの手法においても,

$\alpha_{score}=0$, すなわち難易度のみで単語を抽出した場合との差があまりないという結果になった. また, 青キーワードの正解率も *MDP* (提案カテゴリ) における正解率の最大値と *FLR* における正解率の最大値は 0.539 で等しくなった. これは, 今回用いた, 難易性と専門性という 2 つの概念に共通する部分が多く存在したためだと考えられる. また, *FLR* は大規模なコーパスで適用すると性能が高まる特徴があるため, 今後は, データセット全体で *FLR* を計算し, 評価する必要がある.

(2) データセット 2 による評価

データセット 2 を用いた結果を表 3 と表 4 に示す.

表 3:全キーワード正解率

α_{score}	*MDP* (読売新聞)	*MDP* (提案カテゴリ)	*FLR*
0	0.564	0.564	0.564
0.1	0.570	0.583	0.566
0.2	0.575	0.584	0.549
0.3	0.566	0.577	0.545
0.4	0.566	0.570	0.536
0.5	0.558	0.558	0.532
0.6	0.551	0.557	0.510
0.7	0.544	0.555	0.486
0.8	0.545	0.549	0.469
0.9	0.529	0.529	0.443
1	0.486	0.488	0.404

表 4:青キーワード正解率

α_{score}	*MDP* (読売新聞)	*MDP* (提案カテゴリ)	*FLR*
0	0.531	0.531	0.531
0.1	0.513	0.531	0.531
0.2	0.531	0.575	0.549
0.3	0.549	0.575	0.549
0.4	0.513	0.575	0.549
0.5	0.504	0.549	0.540
0.6	0.504	0.540	0.496
0.7	0.496	0.549	0.504
0.8	0.496	0.549	0.531
0.9	0.478	0.531	0.487
1	0.442	0.460	0.407

表 3 より, *MDP* (読売新聞カテゴリ) における正解率の最大値は 0.575, *MDP* (提案カテゴリ) の最大値は 0.584, *FLR* の最大値は, 0.566 となり, データセット 1 の結果と比較して, 正解率が向上している. 特に, *MDP*(提案カテゴリ)では 5.1 ポイント向上しており, 重要な単語の抽出率が高まっていることがわかる. これは, データセット 1 には, 社会以外に総合などの, 様々な分野にまたがる記事が存在しているのに対して, データセット 2 では, そのような記事を省き, 社会の記事のみが存在しているため, 専門度の精度が高まったことが理由として考えられる. さらに, *MDP* の保有カテゴリに「社会」が存在しているのも理由の一つだと推測できる.

表 4 より, 青キーワードについても全体的に正解率が向上していることが分かる. また, データセット 1 における青キーワードの正解率は *MDP*(提案カテゴリ)の最大値と *FLR* 最大値が等しいが, データセット 2 では, *MDP* (提案カテゴリ) は最大値が 0.575, *FLR* では 0.549 となり, *MDP* (提案カテゴリ) の性能が向上した.

今回の評価では, データセット 2 で向上したものの, 全体的に正解率があまり高くない結果を示している. これは, 前述した理由に加え, 正解データの質も影響していると考えられる. 正解データとして利用した, よみうり博士のアイデアノートでは, 人手で単語に説明を付与しているため, 説明を行う単語の判断基準に揺れの発生が確認された. そこで, 今後は, 小学校の教員や児童に協力を依頼し, より信頼性の高い評価を行う必要がある.

6. おわりに

本稿では, 専門度として *MDP* と *FLR* を, 難易度として, 親密度と抽象度を用いた重要語抽出手法を検討した. 評価の結果, 専門度を計算する場合は, *FLR* よりも *MDP* の方が若干良い結果を与えることが分かった. しかし, 難易度と専門度の正解率の差は僅かであった. 今後は, 難易度と専門度に加え, 主題との関連度を利用した重要度算出手法を検討する.

謝辞

本研究の一部は, 文部科学省科学研究費補助金(若手研究(B) 22700813)の助成を受けて実施した.

参考文献

[NIE 12]
http://nie.jp

[読売 12] よみうり博士のアイデアノート:
http://www.yomiuri.co.jp/nie/note/

[藤沢 12] 藤沢祐輔, 相原慎太郎, 安藤一秋:Web 一般新聞記事を子供向けに言い換える知識の構築, 言語処理学会第 18 回年次大会,pp.751-754 (2012)

[坪井 11] 坪井賢泰, 藤沢祐輔, 小林健, 安藤一秋:小学生を対象とした Web 新聞記事のフィルタリング", 教育システム情報学会第 36 回全国大会, pp.120-121 (2011)

[天野 08] 天野成昭, 笠原要, 近藤公久:日本語の語彙特性第 4 期, 三省堂, 2008.

[池原 99] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦:日本語語彙大系 CD-ROM 版, 岩波書店 (1999)

[中川 03] 中川裕志, 森紘彰, 湯本辰則:出現頻度と連接頻度に基づく専門用語抽出", 自然言語処理, Vol.10, No.1, pp.27-45 (2003)

[久保 10] 久保順子, 辻慶太, 杉本重雄, "異なる学問分野のコーパスを利用した専門用語抽出手法の提案:情報知識学会誌 , Vol. 20 , No. 1 pp.15-31 (2010)

[Rennie 03] Rennie.J.D.M, Shih .L, Teevan .J, and Karger .D.R.: Tackling the Poor Assumptions of Naive Bayes Text Classification", ICML2003, pp. 616-623 (2003)