Webニュース読解支援システムのための重要語抽出の検討

河村 宗一郎[*1]　　安藤 一秋[*2]

[*1]香川大学大学院工学研究科　[*2]香川大学創造工学部

s17g458@stu.kagawa-u.acjp　ando@eng.kagawa-u.ac.jp

1 はじめに

80年代後半より，初等教育機関を中心に，新聞を教材として活用する教育NIE（Newspaper in Education）が行われている[1]．NIEの実践校からは，児童の読解力や表現力の向上，社会への関心が高まるといった効果が報告されている．一方で，新聞記事は児童を対象として書かれておらず，単語や表現が難しいため，児童は新聞記事を読んでも内容を理解できないという問題もある．

本研究では，新聞記事の難しい単語や，記事の主題や教科書の内容に関係する語（以降，重要語と呼ぶ）に対して補足説明を付与する読解支援システムの構築を目的とする．本稿では，新聞記事の段落構造と段落内構造，および単語の専門性に注目して，重要語を抽出する手法について提案する．

2 既存のニュース読解支援

現在，小学生が活用できる読解支援としては，NHKによる「NEWS WEB EASY」がある[2]．これは，図1のように，専門の記者により平易化された記事に対して，ルビと難しい単語への説明文が付与されたニュースサイトである．

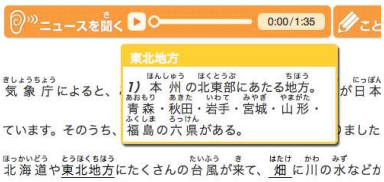

図1 NEWS WEB EASYの記事掲載例

このサービスの記事は1日あたり5本であり，児童が興味に即した記事を選ぶには数が少ない．また全国ニュースや国際問題などの大きな話題を中心として語られるため，地域学習に供するには不向きである．

3 小学生のためのWebニュース読解支援システム

本研究では，新聞記事の難単語や重要語に対して補足説明を付与する読解支援システムの構築を進めている．難単語は，小学生にとって難しい単語である．対象とする学年でまだ習っていない漢字・読み・熟語などが該当する．重要語は，記事の主題を示す単語，もしくは社会科の教科書に出現するような教育的に重要な単語の2つを指すが，本稿では前者についてのみ述べる．

本稿では，新聞記事の構造による単語の重要性，単語の専門性，出現頻度を用いた重要語抽出手法を提案する．我々の研究[3]では，新聞記事の構造が重要語抽出に効果的に作用することを示した．しかし，新聞記事という大きなくくりで専門度を算出していたため，専門度の抽出精度に課題があった．また，抽出辞書の作成に使用した記事データと評価に使用した記事データの年代がかけ離れており，評価の正当性にも問題があった．本稿では，専門度を記事カテゴリごとに算出し，またカテゴリ間の出現頻度も考慮して重要語抽出を行う手法を提案する．

4 抽出候補語の生成

抽出候補語は，単名詞と，単名詞から構成される複合名詞からなる．新聞記事を形態素解析し，次のルールに従って名詞を結合し，単語wを生成する．形態素解析器としてSudachiを用い，ルールの品詞体系はUnidicに準拠する．

1) 抽出対象は名詞/形状詞とする．接頭辞と接尾辞がこれらに連接している場合は，合わせて抽出する．
2) 名詞のうち，第2階層が「助動詞語幹」，または第3階層が「副詞可能」の単語は抽出しない．
3) 形状詞のうち，第2階層が「助動詞語幹」の単語は抽出しない．
4) 接尾辞のうち，第2階層が「動詞的」「形容詞的」または第3階層が「副詞可能」「助数詞」の単語は抽出しない．
5) 2)から4)に当てはまらない名詞/形状詞/接頭辞/接尾辞が連続する場合は，これを結合して1つの複合語として抽出する．
6) 5)のうち，表層系「同」で始まる複合語は抽出しない．
7) 5)のうち，第2階層が「数詞」で始まる複合語は抽出しない．

5 単語の抽出指標

単語の抽出指標として，小林らの手法[4]を参考に，重要度と専門度を定義する．

連絡先：河村宗一郎，香川大学大学院工学研究科，安藤研究室
〒761-0396 香川県高松市林町2217-20

5.1 重要度

重要度は，単語が記事の主題に関係することを表す指標で，文書中の出現位置を利用して求める．

5.1.1 段落構造と段落内構造

新聞記事は見出しやリード文に記事の主題に関する語が利用される[5]．記事に複数の段落がある場合は，各段落内の上位の文についても同様のことがいえる．

見出し，リード文，段落から構成される文書構造を段落構造と定義し，ある単語wが属する段落の段落スコアを$PScore(w)$と呼ぶ．最下段に出現する単語の$PScore(w)$は1，その上の段落の単語は2...とし，見出しの単語は最大値n（nは対象記事の段落数）とする．

同様に，段落内の文から構成される文書構造を段落内構造と定義し，ある単語wが属する文の段落内スコアを$SScore(w)$と呼ぶ．文の数が最も多い段落の文の数を最大値mとし，各段落内の最上段の文に属する単語wの$SScore(w)$はm，その次の文に属する単語はm-1...とする．

例として，段落数が5の記事において，段落2が3文で構成され，最も文数が多い段落とすると，各段落に属する単語の$PScore(w)$と$SScore(w)$は図2のように付与される．

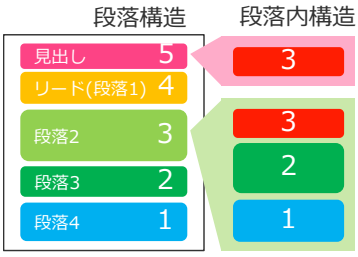

図2 $PScore(w)$と$SScore(w)$の付与例

なお，段落を横断して同じ単語が複数ある場合は，最も高いスコアを$PScore(w)$とする．$SScore(w)$についても同様である．

5.2 専門度

専門度は，単語が専門用語であることを表す指標で，カテゴリ内における単語の連接頻度と，カテゴリ間における単語の出現頻度の差を用いて求める．

5.2.1 カテゴリ内からの専門度算出

カテゴリ内の重要語を抽出する方法として，FLRを使用する．FLRは，中川らが提案した，分野コーパスから専門用語を抽出する手法[6]である．それ以上分割することができない名詞を単名詞，複数の単名詞が連接する名詞を複合名詞と呼ぶとき，複合名詞とこれを構成する単名詞には関連がある．また，ある単名詞が専門分野において重要な概念を表すとき，著者はしばしばその単名詞を含む新しい複合名詞を作る．これらの特徴をもとに，分野コーパス中の単語の連接頻度を用いて専門用語を抽出できる．

本稿では，各記事カテゴリ内の記事集合を分野コーパスと見なし，FLRを適用する．ある単語wのFLR値$FLR(w)$は，解析対象の記事におけるwの出現頻度$f(w)$と，wを構成する各単名詞の連接頻度の相乗平均$LR(w)$を用いて，次の式1で表される．

$$FLR(w) = f(w) \times LR(w) \qquad (式1)$$

また，FLR法とカテゴリ内のIDFを併用することにより，複合名詞を作りやすい単語のうち，一般的な単語の値を引き下げ，専門的な単語の値を引き上げる効果が期待できる．

5.2.2 カテゴリ間からの専門度算出

カテゴリ間からの専門度算出の手法として，MDPを用いる．MDPは，久保らが提案した，「対象分野コーパスと他分野コーパスでの候補用語の出現比率を考慮」した専門用語抽出手法[7]である．

本稿では，対象分野コーパスを解析対象の記事カテゴリ，それ以外のカテゴリを他分野コーパスとして式2のようにMDPを適用する．

$$MDP(w) = minZ_i \quad 1 \leqq i \leqq N \qquad (式2)$$

ただし，
w: 候補語
N: 他カテゴリの総記事数
W_0: 対象カテゴリの候補語の総数
W_i: i番目の他カテゴリの候補語の総数
$f_0(w)$: 候補語Tの対象カテゴリでの出現頻度
$f_i(w)$: i番目の他カテゴリでのTの出現頻度

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

$$\pi_i(w) = \frac{f_0(w)+f_i(w)}{W_0 + W_i}$$

である．

5.3 重要語のスコア計算

単語wの重要度$FLR(w)$，専門度$MDP(w)$の値を使用し，重要語のスコア$score(w)$を計算する．なお，各値はそれぞれ0.0から1.0の値として正規化しておくものとする．段落構造のスコア$spe(w)$を式3とする．

$$spe(w) = x \times PScore(w) + y \times SScore(w) \qquad (式3)$$

ただしx，yは重みである．

カテゴリ内でのIDFを$IDF(w)$とすると，$score(w)$は式4と表せる．

$$score(w) = \alpha \times MDP(w) + \beta \times FLR(w) \times IDF(w) + \gamma \times spe(w) \quad (式4)$$

ただしα，β，γは重みである．

6 評価

6.1 正解データ作成

クローリングによって収集した，2018年1月から11月の読売新聞社のニュース記事から，政治/環境/科学・ＩＴの3カテゴリの約2,000記事を抽出し，これを使用して重要度，専門度の辞書を作成した．このうち12記事をランダムに抽出し，システムで抽出候補語を生成した．この単語群に対し，人手で重要度を付与して正解データを作成した．

重要度の付与は学部生3名の手によって行った．ある単語についてどの程度重要であるかを，それぞれ 重要語/準重要語/重要ではない の3段階で問い，得られた重要度を1.0/0.5/0として合計し，最大3.0の重要度スコアを付与した．このうち，1.0以上のスコアが付与された単語を重要語とした．

6.2 評価方法

システムは，重要度と専門度から，式3および式4を使用して重要語のスコア$score(w)$を計算する．式3の重みは，我々の先行研究[3]より，経験的にx=0.4，y=0.6とする．式4の重みは，仮に$\alpha=\beta=\gamma=1$とする．

$score(w)$を基に重要語を降順にソートし，上位から正解データと同数の単語を抽出して正解データと照合し，F値を求める．また，段落数と同数の単語を抽出し，適合率・再現率・F値を求める．

6.3 結果と考察

重要語のスコアに基づき，上位から正解データと同数の単語を抽出したときの結果を表1に，段落数と同数の単語を抽出したときの結果を表2にそれぞれ示す．

表1では，F値の平均は0.49であり，抽出結果の約半数は重要語が含まれているといえる．カテゴリで見ると，科学・ITカテゴリのF値の平均が他と比較して小さいが，現在の評価データは各カテゴリ4記事であるため，カテゴリ間で抽出精度の差が存在するかデータを増やして確認する必要がある．

実際に重要語抽出を行う際は，重要語として抽出する単語の数も考慮しなければならない．そのため，段落数や文字数などに応じたしきい値を設定する必要があり，段落数に基づいた表2の結果の方がシステムの実使用に近い値と言える．平均値を見ると，適合率よりも再現率の方が値が大きくなった．また，表1の結果と比較すると，F値は6ポイント低くなった．この傾向についてもデータを増やして確認する必要がある．また，抽出する単語数についても検討する必要がある．

表1 正解データと同数の単語を抽出した場合

記事番号	カテゴリ	重要語の数	F値	カテゴリ平均
314	政治	10	0.70	
341	政治	5	0.60	
409	政治	5	0.50	
438	政治	3	0.33	0.53
247	科学・IT	4	0.50	
337	科学・IT	3	0.33	
355	科学・IT	5	0.40	
1706	科学・IT	6	0.50	0.43
212	環境	7	0.57	
288	環境	5	0.60	
1259	環境	3	0.33	
1398	環境	6	0.50	0.50
F値平均			0.49	

表2 段落数と同数の単語を抽出した場合

記事番号	カテゴリ	段落数	適合率	再現率	F値	F値平均
314	政治	5	0.60	0.30	0.40	
341	政治	6	0.50	0.60	0.55	
409	政治	5	0.60	0.50	0.55	
438	政治	5	0.20	0.33	0.25	0.44
247	科学・IT	5	0.40	0.50	0.44	
337	科学・IT	5	0.20	0.33	0.25	
355	科学・IT	8	0.25	0.40	0.31	
1706	科学・IT	5	0.60	0.50	0.55	0.39
212	環境	8	0.50	0.57	0.53	
288	環境	7	0.57	0.80	0.67	
1259	環境	5	0.20	0.33	0.25	
1398	環境	9	0.33	0.50	0.40	0.46
平均			0.41	0.47	0.43	

6.4 今後の課題

抽出候補語の生成部分では，品詞を使ったルールベースの操作では生成できない単語が存在する．偶然に名詞や形状詞が連続した場合や，1語であるべき単語に助詞「の」が含まれる単語は，現状のルールでは抽出できない．また，新聞記事では1度出現した単語は省略されることがあるが，現状のルールでは別語として認識される．

重要度の計算部分では，仮に上位の段落/文から線形にスコアを付与しているが，スコアの付与方法を検討する必要がある．抽出部分では，重要語の抽出数を検討する必要がある．

また，全体として，重要語を含まないと考えられる記事や，教育的に重要な単語を含む記事を処理できるようにする必要がある．

7 終わりに

本稿では，新聞記事の文書構造に加え，カテゴリ内およ

びカテゴリ間の出現頻度を用いて重要語を抽出する手法を提案した．評価実験により，F値は0.43となり，手法の改善および，テストデータを増やした追加実験が必要である．

今後は，抽出候補語の生成ルールの洗練，段落構造/段落内構造の使用率や重みの調査，重要語の抽出数の検討，また小学生向けのシステムとするための教育的要素を取り入れた用語抽出について検討する．

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] 教育に新聞を(http://nie.jp)：日本新聞協会.
(アクセス日：2017年3月15日)

[2] NEWS WEB EASY
(http://www3.nhk.or.jp/news/easy/)：NHK.
(アクセス日：2017年3月15日)

[3] 河村宗一郎，安藤一秋：小学生を対象としたWebニュース読解支援システムのための重要語抽出手法の検討， 2017年度 人工知能学会全国大会 大会論文集，1J1-5, 2017.

[4] 小林健，安藤一秋：小学生を対象とした新聞読解支援のための説明語抽出手法：研究報告コンピュータと教育（CE） 2013-CE-119(17), 1-6, 2013-03-08，2013.

[5] 記者ハンドブック第9版 新聞用字用語集，共同通信社，2001.

[6] 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.

[7] 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.