1J1-5

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

A Keyword Extraction Method for Web News Reading Support System for Elementary School Students

河村 宗一郎[*1]　　安藤 一秋[*2]

Soichiro Kawamura　　Kazuaki Ando

[*1]香川大学大学院工学研究科　[*2]香川大学工学部

Graduate school of Engineering, Kagawa University　　Faculty of Engineering, Kagawa University

NIE (Newspaper In Education) was started at Elementary Schools in the late 1980s. The education uses newspapers as teaching tools. NIE improves children's reading comprehension, and generates various interests in the society. However, it is difficult for elementary school students to read and understand difficult Chinese characters, keywords and the contents of newspaper articles. The aim of our research is to construct a Web news reading support system. This paper considers a keyword extraction method focused on the structure of newspaper articles.

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[NIE 17]．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検討する．

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[NHK 17]．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．

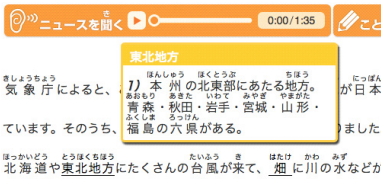

図1 NEWS WEB EASYの記事掲載例

しかし，このサイトでは1日あたり5記事しか掲載されないため，小学生が記事を選ぶには数が少ない．また，このサービスは，留学生などの日本語が苦手な人を対象としたものであり，読解力や表現力の向上を狙うNIEとは目的が異なる．

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．小林らは，難単語と重要語を合わせた，説明が必要な単語「説明語」を定義し，これを自動抽出する手法を提案した[小林 13]．小林らは，説明語は難単語と重要語から成ると定義した．それぞれの指標として難易度，重要度，また両者に関係する専門度を定義し，これらの値を使って難単語，重要語を抽出するものであった．

小林らは，記事の見出しとリード文では重要な単語が使われやすいことを用い，見出しとリード文に含まれる単語と同じ単語を含む文もまた重要であると仮定した．単語をノードとして文を結びつけて重要度の木構造を作成し，ルートと各ノードとの距離を重要語のスコアとして用いた．

しかし，小林らの研究では，段落構造が重要度の算出に与える影響について検証されていない．本稿では，小林らの研究で利用されていた段落構造に加え，段落内構造にも着目し，重要度の算出にこれらが与える影響について検証する．

4. 段落と段落内構造を利用した重要語抽出の検討

重要語抽出における文書構造の有用性について検証するため，ベース手法と文書構造の組み合わせによる重要語抽出を考える．

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

4.1 ベース手法

ベース手法として，TF×IDF，FLR，FLR×IDFの3手法を考える．TF×IDFは統計的な重み付け指標の代表である．FLRは，専門語抽出の手法である．本稿では，新たな指標として，FLRによる専門度スコアに対し，IDFの特性を組み入れたFLR×IDFをベース手法に取り入れる．

以下，TF，IDF，FLRについて，それぞれの考え方を説明する．

(1) TF

TFは，文書中の単語の出現回数である．文書中に頻繁に出現する単語に対して大きな値を与える．

(2) IDF

IDFは，文書集合内である単語が使われている文書数が少ないとき，その単語は文書を特徴付ける単語であるという考え方である．つまり，文書間を横断して頻出する単語は，より一般的な単語であり，重要語にはなりにくいといえる．

ある単語wを含む文書数（文書頻度）をdf_w，文書の総数をNとするとき，wのIDF値は式1で表される．

$$idf(w) = log_2 \frac{N}{df_w} + 1 \qquad (式1)$$

本稿ではベース手法のひとつとしてTF×IDFを採用する．

(3) FLR

FLR法は，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である[中川 03]．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，中川らは，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する手法を提案した．

本稿では，新聞記事集合を分野コーパスと見なし，FLR法を適用する．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2で表される．

$$FLR(w) = f(w) \times LR(w) \qquad (式2)$$

FLR法と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．本稿では，ベース手法のひとつに，FLR×IDFを採用する．

4.2 新聞記事の文書構造

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傾向がある[共同通信社 01]．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．このとき，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，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する．

本稿では，段落構造の指標を段落スコア*PScore*と呼び，段落の最下段を1，その上の段落を2...見出しを最大値n（nは対象記事の段落数）とする．また，段落内構造の指標を段落内スコア*SScore*と呼び，文の数が最も多い段落の文の数を最大値mとして正規化する．つまり，各段落内の最上段の文をm，その次の文を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，*PScore*と*SScore*は図2のように付与される．

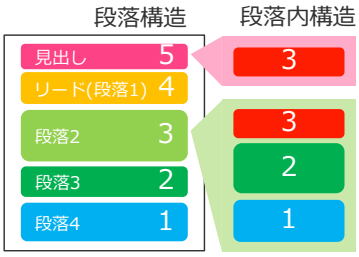

図2 *PScore*と*SScore*の付与例

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*とする．*SScore*についても同様である．

5. 評価実験

重要語抽出において，新聞記事の文書構造の有用性を確認するための評価実験を行う．

ベース手法として，4.で説明したTF×IDF，FLR，FLR×IDFの3つを利用する．これら3手法と文書構造を組み合わせた手法を比較することで，重要語抽出における文書構造の有用性について考察する．

5.1 評価手法

次に示す4つの実験を行い，抽出性能を比較する．

(1) ベース手法のみ

ベース手法として，文書構造を使わずに抽出実験を行う．以下の3つの重要度を用いる．

- TF×IDF
- FLR
- FLR×IDF

(2) 検討1

文書構造として段落構造*PScore*のみをベース手法と組み合わせる．式は次の通りである．

- $weight \times TF \times IDF + (1 - weight) \times PScore$
- $weight \times FLR + (1 - weight) \times PScore$
- $weight \times FLR \times IDF + (1 - weight) \times PScore$

ベース手法と段落構造は，最大値が1となるようにそれぞれ正規化する．重み$weight(0 \leqq weight \leqq 1)$を設定し，重みを

0.0から1.0まで0.1刻みで増加させて11回の抽出実験を行う．

(3) 検討2

文書構造として段落内構造*SScore*のみをベース手法と組み合わせる．式は次の通りである．

- *weight*×TF×IDF+(1−*weight*)×*SScore*
- *weight*×FLR+(1−*weight*)×*SScore*
- *weight*×FLR×IDF+(1−*weight*)×*SScore*

ベース手法と段落内構造は，最大値が1となるようにそれぞれ正規化する．重み*weight*(0≦*weight*≦1)を設定し，重みを0.0から1.0まで0.1刻みで増加させて11回の抽出実験を行う．

(4) 検討3

文書構造として段落構造*PScore*と段落内構造*SScore*の両方をベース手法と組み合わせる．式は次の通りである．

- *weight*2×TF×IDF
  +(1−*weight*2)×(*weight*1×*PScore*+(1−*weight*1)×*SScore*)
- *weight*2×FLR
  +(1−*weight*2)×(*weight*1×*PScore*+(1−*weight*1)×*SScore*)
- *weight*2×FLR×IDF
  +(1−*weight*2)×(*weight*1×*PScore*+(1−*weight*1)×*SScore*)

ベース手法と段落構造，段落内構造は，最大値が1となるようにそれぞれ正規化する．重み*weight*1(0≦*weight*1≦1)，*weight*2(0≦*weight*2≦1)を設定し，重みをそれぞれ0.0から1.0まで0.1刻みで増加させて11回×11回の抽出実験を行う．

5.2 評価用データと評価方法

(1) データの作成

LR辞書，IDF辞書は，1年分の新聞記事と，よみうり博士のアイデアノート[読売新聞社 12]から，2012年の46記事を使って作成する．これには，評価対象の15記事を含む．

(2) 評価方法

正解データは，小林らが先行研究で行ったアンケート結果[小林 13]を使用する．このアンケートは学部生4人に対して行ったもので，よみうり博士のアイデアノートから抽出した評価対象の15記事について，個人が重要と考える単語に0，0.5，1の3段階のスコアが付与されている．本稿では，4人のスコアの合計が2.0以上の単語を正解とする．

抽出した重要語のスコアに上位から順位をつけ，正解データと同数の順位までを取り出す．例として，正解データが10件で，抽出されたデータに10位が3件ある場合，12個の単語が正解として抽出される．

この条件で算出されるF値を正解率として使用する．

5.3 結果

5.1で示した，ベース手法と3つの検討に基づく実験結果を，表1から表5に示す．太字で示したものは最良値である．

表1に示すように，文書構造を使わない場合は，統計的手法であるTF×IDFが最も高い正解率を示した．しかし，3手法の正解率の差は大きくない．

表2に示すように，段落構造のみを使用した場合の最良値は，TF×IDFとの組み合わせではベース手法と同値，FLR，FLR×IDFではベース手法の正解率を2ポイント程度上回り，検討1の正解率よりも数ポイント高かった．

表3に示すように，段落内構造のみを使用した場合の最良値は，TF×IDF，FLRとの組み合わせではベース手法を4ポイント程度上回った．FLR×IDFについては，ベース手法を上回ったものの，検討1の結果とほぼ同じ値となった．

表1 ベース手法の実験結果

	TF×IDF	FLR	FLR×IDF
BASE	0.250	0.225	0.240

表2 検討1 の実験結果

weight	TF×IDF	FLR	FLR×IDF
0.0	0.110	0.110	0.110
0.1	0.144	0.118	0.125
0.2	0.144	0.118	0.112
0.3	0.135	0.110	0.147
0.4	0.150	0.146	0.212
0.5	0.223	0.211	0.208
0.6	0.240	0.244	0.258
0.7	0.237	0.238	**0.267**
0.8	0.248	**0.251**	0.241
0.9	0.244	0.205	0.223
1.0	**0.250**	0.225	0.240

表3 検討2の実験結果

weight	TF×IDF	FLR	FLR×IDF
0.0	0.157	0.157	0.157
0.1	0.231	0.240	0.241
0.2	0.231	0.240	0.241
0.3	0.231	0.252	0.255
0.4	0.214	0.252	0.255
0.5	0.251	**0.265**	0.259
0.6	0.251	0.257	0.261
0.7	0.271	0.240	**0.266**
0.8	**0.289**	0.238	0.243
0.9	0.266	0.215	0.242
1.0	0.250	0.225	0.240

表4は，検討3において，TF×IDFについての最良値を含む結果を示したものである．このときのweight2の値は0.8である．また表5は，検討3において，FLR，FLR×IDFについての最良値を含む結果を示したものである．このときのweight2の値は0.5である．

表4 検討3 TF×IDFの実験結果

weight2	weight1	TF×IDF
0.8	0.0	**0.289**
	0.1	0.253
	0.2	0.254
	0.3	0.249
	0.4	0.249
	0.5	0.257
	0.6	0.249
	0.7	0.249
	0.8	0.240
	0.9	0.232
	1.0	0.248

表5 検討3 FLR，FLR×IDFの実験結果

weight2	weight1	FLR	FLR×IDF
0.5	0.0	**0.265**	0.259
	0.1	0.253	0.271
	0.2	0.253	0.279
	0.3	0.262	0.292
	0.4	0.219	**0.314**
	0.5	0.219	0.237
	0.6	0.230	0.232
	0.7	0.242	0.223
	0.8	0.231	0.228
	0.9	0.231	0.219
	1.0	0.211	0.208

表4で，段落構造と段落内構造を組み合わせて使用した場合の最良値は表3の最良値と同値となり，段落内構造のみが有効であるという結果となった．また，表5のFLRについても同様であった．FLR×IDFについては，ベース手法を7ポイント程度，検討1，2の結果を2ポイント程度上回り，段落構造と段落内構造の組み合わせが有効であるという結果となった．

5.4 考察

以上の3つの実験より，TF×IDF，FLRについては，段落内構造との組み合わせが特に有効であることを確認した．また，FLR×IDFについては，段落構造と段落内構造の併用が有効であることを確認した．このような違いはあるが，いずれの場合も専門度の算出と文書構造を組み合わせた方が，ベース手法と比べてよい結果を得ることができた．これより，新聞記事からの重要語抽出において，文書構造が有用に働くことがわかった．

今回の実験では，各記事内の段落数の差については考慮しておらず，段落スコア，段落内スコアの減少率は線形と仮定している．また，段落内スコアは，すべての段落の中で最大の文数を使って正規化しており，段落内構造には段落構造の性質も含まれている．このような理由から，手法によって有効な文書構造の要素が異なった可能性が考えられる．今後は，段落数の差も考慮しながら，文書構造の各要素の有効性について記事単位での調査を行う必要がある．また段落スコア，段落内スコアの減少率についても検討する必要がある．

正解率については，文書構造を用いることによりベース手法から上昇することが確認できたが，現時点での最良値は3割程度にとどまっている．これは，解析対象の記事が2012年であるのに対し，各種辞書データの作成に使用した記事が10年以上前のものであるため，これ以降に出現した単語についてIDF値が正確でないことが原因の1つとして考えられる．

今後は，解析対象記事と辞書作成用のデータの年代をそろえたデータを用意し，再実験を行って正解率が向上するかを確認する必要がある．

6. おわりに

本稿では，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について述べた．新聞記事からの重要語抽出について，3つのベース手法を考え，文書構造を用いることにより正解率が向上することを確認した．

今後は，各記事について文書構造の要素の有効性について調査するとともに，正解率を向上させるための辞書作成についても検討する．

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

[NIE 17] 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)

[NHK 17] NEWS WEB EASY (http://www3.nhk.or.jp/news/easy/)：NHK．(アクセス日：2017年3月15日)

[小林 13] 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.

[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.

[共同通信社 01] 記者ハンドブック第9版 新聞用字用語集，共同通信社，2001.

[読売新聞社 12] よみうり博士のアイデアノート (http://www.heu-le.net/yomi3/top.html)：読売新聞．(アクセス：2012年．2017年現在サービス終了．)