

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

小林 健^{†1} 安藤一秋^{†2}

近年、小学校をはじめとする教育機関で、新聞を活用する教育 NIE (Newspaper In Education) が注目されている。しかし、小学生にとって新聞の内容は難しく、読んでも理解できない場合がある。そのため、NIE を効果的に実施できない問題がある。そこで本研究では、Web 新聞読解支援システムの構築を目的とする。本稿では、説明すべき単語 (説明語) の抽出手法を提案する。

A Keyword Extraction Method for Newspaper Reading Support System for Elementary School Students

KEN KOBAYASHI^{†1} KAZUAKI ANDO^{†2}

Recently, NIE (Newspaper In Education) has become increasingly active at elementary schools. However, there are Kanji characters or terms which students cannot read and understand in newspapers. In order to increase the effectiveness of NIE, we aim to construct a Web newspaper reading support system for elementary students. This paper proposes a method for extracting keywords which should be explained for elementary school students from newspapers.

1. はじめに

近年、小学校などの教育機関で新聞を教材として活用する教育 (NIE : Newspaper in Education) [1]が注目されている。特に、平成 23 年度から実施されている新教育指導要領 [2]では、新聞を授業の中で利用することが明示され、その重要性は益々高まっている。

NIE の効果としては、NIE の実態調査結果報告 [3]の中で、子どもの読解力や表現力の向上、社会に対する関心の高まりや、日常における新聞の閲読機会の向上などが報告されている。

様々な効果が期待できる NIE であるが、NIE の実態調査や我々が行ったアンケートによって、以下の問題と意見が明らかになった。

- (1) 新聞は難しい表現や単語が多くわかりにくい。
- (2) 多くの小学生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) 補足資料として新聞記事の内容に関連する写真やグラフが欲しい
- (5) 自動の読解力に個人差が存在している。

以上を解決・実現することで、NIE の効果を高めることができると考えられる。そこで我々は、(1)を解決するため、新聞中の難しい表現を分かりやすい表現に言い換える研究 [4]や、(2)と(3)を解決するために、記事推薦に関する研究 [5]に取り組んでいる。本研究では、(4)や(5)を解決するため、小学生を対象とした読解支援システムの構築を目指してい

る。本稿では、読解支援システムの概要と小学生に説明すべき単語 (説明語) の抽出手法について提案する。

2. 読解支援システム

2.1 よみうり博士のアイデアノート

小学生を対象とした新聞読解支援システムとして、読売新聞社が Web 上で公開している、「よみうり博士のアイデアノート [6]」が存在している。このシステムは、以下の機能を持つ。

- (1) 難しい用語の説明
- (2) 調べ学習の起点となる単語の提示
- (3) 記事の内容に関連するグラフやデータの提示

(1)と(2)について詳しく説明する。アイデアノートには、新聞を子ども向けに言い換えた文と原文へのリンクが存在しており、通常の新聞より読みやすくなっている。語彙の少ない小学生に配慮するため、言い換えられた文の中でも、難しい単語は「赤」で強調表示され、簡単な説明が付与されている。また、調べることでより記事の内容理解が増す、調べ学習の起点となる単語は「青」で強調表示されている。この単語をクリックすると、関連情報が掲載された Web ページへのリンクや、教科書の関連部分が記述されており、新聞の理解をより深めることができる。

しかし、これらの支援はすべて人手で行われているため、更新速度に問題があるだけでなく、すべての一般記事に対応していないので、利用したい記事が必ず存在している保証はない。

2.2 NEWS WEB EASY

NEWS WEB EASY [7]は、人にやさしい放送提供を目的に、小中学生や日本居住の外国人に、わかりやすいことばでニ

^{†1} 香川大学大学院工学研究科
Kagawa University, Graduate School of Engineering.
^{†2} 香川大学工学部
Kagawa University, Faculty of Engineering

ユースを伝えるサービスである。支援機能として以下を提供している。

- (1) 地名、組織名、人名を色で分けて表示
- (2) 難しい単語に簡単な説明を付与

NEWS WEB EASY では、地名を「青」、人名を「紫」、組織名を「水色」で表示することで、固有表現と一般的な単語を区別することで閲読を支援している。また、難しい単語には、小学国語辞典の説明を付与し、マウスオーバーで閲覧できる。

これらの支援を利用することで、難しい単語に躓くことなく記事を読むことができる。しかし、記事を深く読み取るための支援は行われていない。また、人手で行われる支援であるため、よみうり博士と同様、記事数の問題もある。

2.3 提案システム

2.2 と 2.3 で述べた既存の読解支援システムは、すべて人手で支援を行なっているため記事数が少なく、自分の読みたい記事に対して支援を受けることができない可能性が高い。そこで我々は、Web 上で公開されている一般新聞記事に対して、以下の支援を自動で行う読解支援システムを構築する。

- (1) 難しい語や重要語に対する自動説明付与
- (2) 記事に関連するデータや写真の提示
- (3) 複数の新聞記事の比較
- (4) 調べ学習支援

(1)~(4)について簡単に説明する。(1)の機能は、記事中の難しい語や主題に関わる重要な語を自動判別し、説明を付与することで読者の読解を支援する。小学生の語彙能力等に合わせて、説明を行う単語数や内容を調整することを考えている。(2)の機能は、記事中の重要語等を利用して、Web 上から関連データや画像などを収集し、読者に提示する機能である。(3)の機能は NIE において、同一内容の記事を異なる新聞社で比較するという実践が行われているが、それを支援するための機能である。また、異なる新聞社の内容を比較することで、別の視点の論調を学習できる。(4)の機能は、NIE が総合学習で利用されやすいということに着目した機能で、新聞記事の内容から調べ学習課題を自動生成する機能[7]の実装を進めている。

(1)~(4)の機能を実装するためには、新聞記事の内容に大きく関わる重要な語や記事中の難しい語を抽出することが必要不可欠である。そこで本稿では、難しい語（難易語）と主題に関わる重要な語（重要語）を説明すべき単語（説明語）と定義し、それらを抽出する手法を提案する。

3. 説明語抽出手法

3.1 説明語抽出の概要

本研究では、よみうり博士のアイデアノートを参考に、説明語を、「難しい語（難易語）」と「文章の主題に関係する単語（重要語）」に分けて考える。これらの単語がもつ特

性をそれぞれ、「難易度」および「重要度」と定義する。また、この双方に関係する指標として「専門性」を考え、その指標として「専門度」を定義する。専門度と重要度、難易度の関係を図に示す。

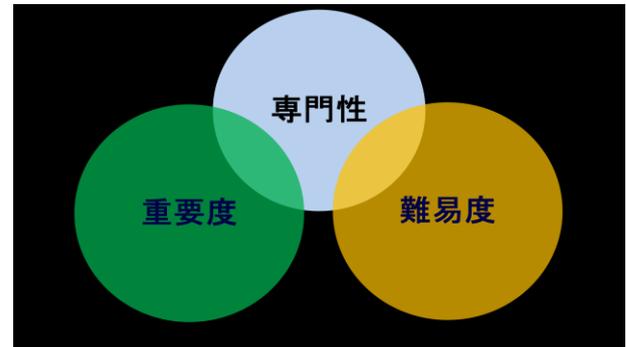


図1 専門度と難易度、重要度の関係

最終的に、難易度と重要度を専門度と組み合わせてスコア計算し、その上位を難易語、重要語として抽出する。

3.2 専門度の算出

専門度とは、単語の専門性を表す指標であり、図1に示すように、重要度と難易度に大きな関係すると考えられる。本稿ではそ、既存の専門用語抽出手法である FLR[8]と MDP[9]を用いて専門度を計算し、難易語と重要語の抽出に利用する。

3.2.1 FLR

FLR とは、単語の接続頻度を利用し、分野コーパスから専門用語を抽出する手法である。接続頻度を利用することで、分野固有の概念への関連強度であるターム性を表現している。

FLR は、以下で計算される。

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

$$FLR(W) = f(W) \times LR(W) \quad (2)$$

ここで、

W : 名詞。

$FL(w)$: w が他の単名詞の右側に接続する頻度。

$FR(w)$: w が他の単名詞の左側に接続する頻度。

w_i : 単語 W に含まれる単名詞

n : 単語 W を構成する単名詞数

$f(W)$: 単語 W の出現頻度

である。

3.2.2 MDP

MDP (Minimum of the Difference between Population Proportions) 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し、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。

$$\text{class_MDP}(W, j) = \min Z_i \quad 1 \leq i \leq N \quad (3)$$

$$\text{MDP}(W) = \max \text{class_MDP}(W, j) \quad 1 \leq j \leq N+1 \quad (4)$$

$$Z_i = \frac{\frac{f_j(W)}{W_j} - \frac{f_i(W)}{W_i}}{\sqrt{\pi_i(W)(1-\pi_i(W))\left(\frac{1}{W_j} + \frac{1}{W_i}\right)}} \quad (5)$$

ここで、

W : 単語

j : 対象分野コーパス

N : 他分野コーパスを構成する分野の数

W_j : 対象分野 j 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

$f_j(W)$: 単語 W の対象コーパス j での出現頻度

$f_i(W)$: i 番目の他分野コーパスにおける単語 W の出現頻度

$$\pi_i(W) = \frac{f_j(W) + f_i(W)}{W_j + W_i} \quad (6)$$

である。

MDP では、異なる分野のコーパスを利用して、スコアを算出する。そのため、どのような分野コーパスを構成するかによって MDP の値も変化する。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。

本稿では、Web 新聞記事として、読売新聞社の Web ページから 2011 年 5 月に収集した 2,891 件の新聞記事とよみうり博士掲載記事を元に分野コーパスを構成する。また、収集した記事を 10 分野に分類する際、コンプリメントナイーブベイズ[10]を利用する。

MDP を利用する際、あらかじめ各単語の MDP が保存された単語辞書を利用する。

3.2.3 専門度の算出

専門度は、 FLR と MDP の値を足しあわせて計算する。これは、 FLR と MDP がそれぞれ異なる視点から専門用語を抽出する手法であり、その視点の双方を取り入れ、専門度を計算するためである。

専門度 $spe(w)$ は以下のように計算する

$$spe(w) = \frac{MDP(w)}{2MAXMDP(D)} + \frac{FLR(w)}{2MAXFLR(D)} \quad (7)$$

ここで、

D : 新聞記事

$MAXMDP(D)$: 記事 D における MDP の最大値

$MAXFLR(D)$: 記事 D における FLR の最大値

である。

3.3 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。

大人も子供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。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可能性が高いと考えられる。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。

3.3.1 親密度

親密度は、日本語の語彙特性[11]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 (けいてい)」に対する単語親密度は、「市場」= 6.188, 「逕庭」= 1.5 と定義されている。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。

本研究では、単語 w の親密度 $sim(w)$ を次式で定義する。

$$sim(w) = 1 - termsim(w) / 7 \quad (8)$$

ここで、 $termsim(w)$ は w の単語親密度であり、7 は単語親密度の最大値である。

3.3.2 抽象度

抽象度とは、単語の意味がもつ広さを数値で表現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[12]において、「経験」と「猛省」のシソーラス上の深さは、それぞれ 4, 9 となる。深さが深い「猛省」の方がより具体的で、難しい単語であると考えられる。

本研究では、単語 w の抽象度 $abs(w)$ を次式で定義する。

$$abs(w) = depth(w) / 13 \quad (9)$$

ここで、 $depth(w)$ はシソーラス上の深さである。13 は日本語語彙体系における深さの最大値である。

3.3.3 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 の難易度 $dif(w)$ を次式で定義する。

$$dif(w) = (1 - wg_1) \frac{sim(w)}{simMAX(D)} + wg_1 \frac{abs(w)}{absMAX(D)} \quad (10)$$

ここで、

$simMAX(D)$: D 内の親密度の最大値

$absMAX(D)$: D 内の抽象度の最大値

wg_1 ($0 \leq wg_1 \leq 1$): 重み

3.4 重要度

文章の主題に関係する単語である重要語が持つ特性値として重要度を定義する。本研究では、新聞記事が対象であるため、記事の構造に着目して、重要度を推定する手法を提案する。

新聞記事は、見出しや1文目に記事の主題に関係する語が利用されやすいという特徴がある[13]。したがって、見出しと1文目を重要度計算に活用することで、単語の重要性を表現する。具体的には、文のスコアとして、含まれる単語の特徴度(*idf*)、タイトルからの距離、共通する単語を通じて繋がっている親の文との繋がりを考慮してスコアを計算していく。

以下に重要語抽出の手順を示す。

- ① 記事 D (N 文, M 段落) を形態素解析後、複合名詞化を行い、見出し、文、段落ごとに登録する。
- ② 見出し及び1文目を0次重要文とし、そこに含まれている単語を1次重要語として登録する。
- ③ n 次重要語を含む文を n 次重要文とし、含まれている n 次重要語以外の単語を $n+1$ 次重要語として登録する。
- ④ 未探索の文の中から新たな n 次重要文が見つければ、③に戻り、さもなければ、⑤の処理を実行する。
- ⑤ 0次重要文 s_0 に対する文重要度 $scoreS$ を計算する。

$$scoreS(s_0) = idf_MAX(s_0) + \alpha \quad (11)$$

ここで、

$idf_MAX(s)$: 文 s に含まれる単語の idf の最大値

α : s_0 中に固有表現を含む場合 1 それ以外 0

である。

- ⑥ 0次重要文以外の n 次重要文 s_i ($1 \leq i \leq N$) に対する文重要度 $scoreS$ を計算する。

$$scoreS(s_i) = \frac{\beta}{n+1} \{idf_MAX(s_i) + parent_score(s_i)\} \quad (12)$$

$$parent_score(s_i) = \sum_p connect(s_i, p) \times scoreS(p) \quad (13)$$

$$connect(s_i, p) = \frac{idf(link)}{\sum_{anc} idf(anc)} \quad (14)$$

ここで、

β : 段落内において一文目である場合 1 それ以外 0

p : s_i に結びつく親の文

$link$: s_i と p を結びつけている単語

anc : p と下位の重要文を結びつけている単語

$idf(w)$: 単語 w の idf 値

である。

- ⑦ 記事 D に含まれる異なり単語 t_j ($1 \leq j \leq L$) の単語重要度 $imp(t_j)$ を計算する。

$$imp(t_j) = \sum_{sen} scoreS(sen) / rank(sen) \quad (15)$$

ここで、

sen : 単語 t を含む n 次重要文

$rank(sen)$: sen が存在する段落位置 k ($1 \leq k \leq M$)

である。

(11)~(15)式についてそれぞれ説明する。(11)式は、見出しと1文目のスコア計算式である。見出しや1文目は、記事を読む上で重要な場合が多いが、その中に特徴的な単語が存在する／しない場合を比べると、重要性が異なると考えられる。そこで、単語の特徴度を示すために、文に含まれる単語の idf の最大値を利用する。また、タイトル中に固有名詞が存在する場合、その固有名詞が記事の主題に大きく関わると仮定し、固有名詞の有無によって、スコアに補正を加える。

(12)式は、 n 次重要文自身の idf と親の文から伝搬したスコア $parent_score$ の和を $n+1$ で除算している。これは、 n が大きい程、タイトルとの関係が小さくなると考えられるため、このような補正を行う。また、各段落の1文目は、その段落で最も述べたい事柄 (topic sentence) が書かれるため、スコアが高くなるように重み β を掛けている。

(13), (14)式は、親から子に伝搬するスコアを計算する式である。具体的には、親から特徴的な単語で子に繋がれる場合、一般的な単語で繋がれる場合より、多くの情報が親から子に伝搬されると仮定し、スコアを伝搬させる。

3.5 説明語の抽出

ここまでで述べた指標を用いて、難易語と重要語を抽出するためのスコアを定義する。

3.5.1 難易語の抽出

難易度と専門度を組み合わせ、記事 D における難易語 w のスコアを以下の式で計算する。

$$score_dif(w) = (1 - wg_2) \frac{dif(w)}{difMAX(D)} + wg_2 \frac{spe(w)}{speMAX(D)} \quad (16)$$

ここで、

$difMAX(D)$: 記事 D における難易度の最大値

$speMAX(D)$: 記事 D における専門度の最大値

wg_2 ($0 \leq wg_2 \leq 1$) : 重み

である。

最終的に、 $score_dif(w)$ の上位を難易語として抽出する。

3.5.2 重要語の抽出

重要度と専門度を組み合わせ、記事 D における重要語 w のスコアを以下の式で計算する。

$$score_imp(w) = (1 - wg_3) \frac{imp(w)}{impMAX(D)} + wg_3 \frac{spe(w)}{speMAX(D)} \quad (17)$$

ここで、

$impMAX(D)$: 記事 D における難易度の最大値

$speMAX(D)$: 記事 D における専門度の最大値

wg_3 ($0 \leq wg_3 \leq 1$) : 重み

である。

最終的に、 $score_dif(w)$ の上位を重要語として抽出する。

4. 評価実験

4.1 実験方法

前述した、(16)式、(17)式の重みを変化させながら、難単語、重要語を抽出し、正解データと比較し、抽出性能を評価する。正解データには、2つのデータセットを用いる。

データセット1は、読売博士のアイデアノート記載の46記事を利用し、その中で、赤で強調表示されている単語を難単語の正解、青で強調表示されている単語を重要語の正解データとする。以下に示す正解率で評価する。

$$\text{正解率} = \frac{\sum_{link} \text{抽出された正解の個数}}{\text{正解単語の個数}} \quad (18)$$

データセット2は、学部生4名に対して行ったアンケートを基に作成した。アンケートの際、読売博士のアイデアノートの社会カテゴリ掲載の記事15件をランダムに選択し、記事中の単語に対し、難しい単語か、主題に関係する重要な語であるかをそれぞれ、0, 0.5, 1の3値で評価してもらった。なお、強く関係する場合を1、少し関係する場合を0.5、関係しない場合に0とする。本実験では、4人の評価値の総和が2.5を超える単語を正解とする。ここで、(10)式の重み Wg_1 は予備実験により0.3とする。また、単語を抽出する個数は、正解となる単語の数とした。

4.2 実験結果(データセット1)

データセット1に対して、難易語の抽出実験を行った結果を表4.1に示す。表4.1を見ると、 Wg_2 が0.1の時、正解率が47.15となっており、専門度と難易度を組み合わせることで抽出性能が向上していることがわかる。この結果より、難易度と専門度に関係性があることが確認できた。また、重みを増やす(専門度の比重を上げる)と正解率は下がっており、難単語抽出には難易度が大きく寄与していることがわかった。しかし、正解率の最大値は47.15と、あまり良い結果ではなかった。これは、現状、小学生にとって難しい単語を完全に表現できていないということが考えられる。また、よみうり博士のアイデアノートにおける赤単語と青単語の定義が揺れていることが、原因の一つとして考えられる。

表4.2に重要語抽出実験を行った結果を示す。表4.1より、 Wg_3 が0.9(重要度と専門度が1:9の比率)で重要語を抽出した時、正解率が21.43で最大となり、正解率はあまり高くないものの、専門度と重要度を組み合わせることの有用性を確認できた。今回の重要語抽出実験で性能が低くなった理由は、よみうり博士のアイデアノートにおける、重要な単語(青単語)の個数が非常に少ないためだと考えられる。よみうり博士における青単語の総数は1記事に2~3個程度で、正解データの個数だけ抽出する実験を行った場合、抽出する単語そのものが少なくなってしまうため、正解率も下がってしまった。それに加え、青単語の基準が曖昧であるということも正解率が低くなった理由として、

考えられる。重要度のみで重要語を抽出した場合($Wg_3=0$)に最も正解率が低くなっているが、これは、現状の手法では、文の重要度を差別化できるが、同一の文に存在する単語の重要度には差ができない。そのため、あまり重要でない単語も、場合によってはスコアが高くなり、正解率を下げる原因になっている。

表 4.1 難単語抽出性能 (データセット1)

Wg_2	正解率
0	46.56
0.1	47.15
0.2	45.78
0.3	44.99
0.4	42.63
0.5	41.85
0.6	41.06
0.7	40.08
0.8	39.10
0.9	38.70
1	37.92

表 4.2 重要語抽出性能 (データセット1)

Wg_3	正解率
0	11.22
0.1	13.27
0.2	15.31
0.3	15.31
0.4	15.31
0.5	17.35
0.6	19.39
0.7	19.39
0.8	19.39
0.9	21.43
1	20.41

4.3 実験結果(データセット2)

データセット2に対して難易語の抽出実験を行った結果を表4.3に示す。表4.3を見ると、 $Wg_2=0.1$ の時、正解率が48.08で最大となっている。この傾向は、データセット1で実験を行った場合と類似した結果となった。

データセット2に対して、重要語の抽出実験を行った結果を表4.4に示す。表4.4より、 $Wg_3=0.7, 0.8, 0.9$ の時、正解率49.15で最大となっており、重要度と専門度を組み合わせることが有用であることが確認できた。また、全体的にデータセット1と比較して、正解率が良くなった。これは、アンケートに際して、主題に関係する単語ということを意識してもらうことで、本手法の評価に適した単語を

得ることができたと考えられる。

表 4.3 難易語抽出性能(データセット 2)

W _{g₂}	正解率
0	44.23
0.1	48.08
0.2	46.15
0.3	42.31
0.4	38.46
0.5	34.62
0.6	30.77
0.7	28.85
0.8	28.85
0.9	26.92
1	23.08

表 4.4 重要語抽出性能 (データセット 2)

W _{g₃}	正解率
0	20.34
0.1	27.12
0.2	27.12
0.3	28.81
0.4	28.81
0.5	32.20
0.6	35.59
0.7	49.15
0.8	49.15
0.9	49.15
1	47.46

5. おわりに

本稿では、新聞読解支援システムの概要について述べ、小学生に対して説明すべき単語として、難易語、重要語を定義した。そして、それらを表現する指標として、難易度、重要度を考え、専門度と組み合わせることで、難易語、重要語を抽出する手法を提案した。その結果、難易度、重要度ともに、専門度と組み合わせることによって、最も高い抽出性能を示した。しかし、抽出性能は難易語、重要語ともに最大で 50%程度であり高くなく、特に重要語抽出に関しては、重要度単独で単語を抽出した際の正解率がかなり低く、特に改良が必要である。

今後は、精度、再現率を用いた抽出性能の評価や、重要語抽出手法の改良を行う。具体的には、現在、重要度算出時、同一の文に含まれる単語は、同じ重要度が割り当てられてしまう。そのため、タイトルに含まれる動詞や係り受け関係等を利用して、単語間の重要度の差別化を図る必要がある。

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

- 1) 教育に新聞を <http://nie.jp>
- 2) 新学習指導要領・生きる力
http://www.mext.go.jp/a_menu/shotou/new-cs/index.htm
- 3) NIE 実践の実態調査結果報告 http://nie.jp/inves/ji1_200807.pdf
- 4) 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抽出, 言語処理学会第 18 回年次大会論文集, p.751-754 (2012).
- 5) 坪井賢泰, 安藤一秋: 小学生向け NIE を対象とした Web 新聞記事の推薦, 情報処理学会研究報告, コンピュータと教育研究会報告, 2013-CE-119 (2013).(発表予定)
- 6) よみうり博士のアイデアノート
<http://www.yomiuri.co.jp/nie/note/>
- 7) NEWS WEB EASY <http://www3.nhk.or.jp/news/easy/>
- 8) 久保恵津子, 小林健, 坪井賢泰, 安藤一秋: 調べ学習課題の自動生成に向けた学習課題の分析, 第 11 回情報科学技術フォーラム講演論文集, 第 4 分冊, pp.301-302 (2012).
- 9) 中川裕志, 森紘彰, 湯本辰則: 出現頻度と接続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)
- 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)
- 11) 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)
- 12) 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.
- 13) 池原悟, 宮崎正弘, 白井論, 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)
- 14) 共同通信社, “記者ハンドブック新聞用字用語集第 9 版”, 2001, 共同通信社