Vol.2015-DBS-162 No.27
2015/11/26

Webニュースからの観点抽出手法の提案

大原 正章[1,a)]　真下 遼[1,b)]　灘本 明代[1,c)]

概要：Web ニュースを閲覧する際，1 つの記事を読んだだけでは内容の重要性を把握できない場合がある．このような時，多くのユーザが記事ページ内でリンクされている関連記事を閲覧すると考えられるが，これら関連記事のほとんどは記事の経過情報を記載したニュースや記事の主題が同じ他のニュースであるため，ユーザが閲覧している記事の重要性を判断する手掛かりにはならない．そこで，閲覧記事の主題に対してライバルに関する記事でかつ閲覧記事の内容と類似したニュースを対立記事として提示し，比較できれば閲覧記事の重要性を理解する手助けになると考えられる．またこの時，記事には複数の観点が存在する場合がある．そこで，本研究では閲覧記事の観点毎の対立記事を抽出する手法について提案する．

キーワード：ニュース，観点，Web，対立記事，主題語

1. はじめに

現在，Web ニュースが普及しており様々なニュースサイトが存在している．Web ニュースはいつでもリアルタイムにニュースを取得することが出来るため，Web ニュースを利用することは最新ニュースを取得する有用な手段の一つであると考えられる．しかしながら，Web ニュースを閲覧する際，1 つの記事を読んだだけでは内容の重要性を把握できない場合がある．例えば，「又吉が芥川賞を受賞した書籍の発行部数が１００万部を突破した」という記事を閲覧した際に，芸人が書いた本の発行部数や芥川賞受賞のほんの発行部数が分からない場合，１００万部の突破がどれほどの偉業であるのかを理解することは困難である．このような場合，一般に Web ニュースのページの下部に掲載されている関連記事を見る場合が多い．しかしながら，関連記事の多くは閲覧している記事の過去に報道された記事である場合や，閲覧記事内に出現しているキーワードに関連する記事である場合がほとんどである．そのため，関連ニュースを閲覧しても，元のニュースの重要性を理解することは困難である場合が多数ある．

そこで我々は，閲覧記事と対立関係にある記事を提示することにより，その記事の重要性を理解することが可能であると考え，閲覧記事の対立関係にある記事を抽出し提示する手法を提案する．本論文では，対立関係にある記事を対立記事と呼ぶ．この時，ニュース記事には記事内で話題の中心となる語である主題があり，主題に対する述語がある．本論文では，記事内の主題となる語を主題語と呼ぶ．そして主題語には複数の観点がある場合が多い．例えば，上記の例では「又吉」が主題であり，発行部数が述語であるとする．この時，この記事の観点は「芸人」や「芥川賞」等が考えられる．対立記事は「芸人」を観点とすると主題のライバルは他の芸人であり，ライバルの記事は「芸人の本の発行部数」に関する記事である．また，「芥川賞」を観点とすると主題のライバルは他の芥川賞を受賞した作家であり，ライバル記事はその作家の発行部数に関する記事になる．このように観点によって対立記事は異なる．この時，「芥川賞」は記事の中に記載されているため，観点として容易に取得することができる．しかしながら「芸人」は記事の中に記載されていない，人々が知識として知っている観点である．そこで我々はニュースの観点として明示的観点と暗黙的観点があると考え，これらの観点を抽出し，その観点毎の対立記事を抽出する手法を提案する．

以下，2 章では関連研究を紹介し，3 章では提案手法について述べる．そして 4 章では実験について述べ，5 章でまとめと今後の課題について述べる．

2. 関連研究

Web ニュース記事の理解を支援することを目的として比較対象を抽出する研究は多数存在する．池田ら [1] は，ニュース記事と blog 記事を対応付けることでニュースの理解を支援する手法を提案している．本研究では，対応付ける対象はニュース記事同士である点が異なる．北山ら [2]

[1] 甲南大学
Konan Uniersity, Kobe, Hyogo 658–8501, Japan
[a)] m1524001@s.konan-u.ac.jp
[b)] hr-x7type@yahoo.co.jp
[c)] nadamoto@konan-u.ac.jp

は，映像ニュースとテキストニュースの比較のための質問生成の提案を行っている． 本研究とはテキストニュースである Web ニュース記事のみを扱っている点で異なるが，記事内の特徴語抽出において，一般に Web ニュース記事では理解に重要なことから先に書かれている点に着目して単語の重要度を決めており，本研究でもニュース記事の本文一段落目と二段落目以降で異なる単語の重要度を付与している．切通ら [3] は，ニュース記事から固有名詞に関する記述の差異に着目し，関連記事の関連度や擁護度など 4 つの尺度によって関連ニュースのランキングを行う手法を提案している．また、主題となる語を tf-idf 法を用いて抽出を行っているのに対し，本研究では，単語の出現位置を考慮している点で異なる．

また，ニュース記事から主題となる特徴語の抽出について，田中ら [4] は，記事内で現れる人物や組織，場所，建造物などのエンティティをニュース記事の話題となる特徴語として複数の語を抽出している．本研究でも主題となる特徴語は人物や組織等の固有名詞であることに着目して抽出を行っているが，単語の出現位置による重要度の付与によって一語のみの特徴語抽出を行っている点で異なる．

ニュース記事における特徴語となる話題語抽出の手法としてトピックモデルを用いた研究では，佐藤ら [5] は，複数のニュース記事において「政治」「スポーツ」「経済」などの分類を行い，更にパラメトリック混合モデルを基にした分類手法を用いることで特徴語である話題語の抽出を行っている．菊池ら [6] は，ニュース記事などの時系列テキスト集合において単語ベクトルの余弦尺度を用いて話題後を抽出している．高橋ら [7] は，世の中の特異な出来事に対して関連する記事が急激に増加する点に着目してダイナミックトピックモデルを用いることでトピック単位のバースト検出による話題語抽出を行っている．本研究では，潜在的トピック配分法（LDA: Latent Dirichlet Akkication）[8] のトピックモデルを用いている点で類似しているが，主題となるニュース記事の特徴語についてはクエリとなる一つの記事から抽出する点で異なる．

さらに，ニュース記事を対象に LDA を用いた研究では，吉田ら [9] は，ニュース記事の記述から株価の取引高を予測するために，同一の話題の記事がまとまるようクラスタリングを行っている．しかしながら，文書の対象を記事のタイトルのみとして表記揺れを考慮した LDA の拡張である Dirichlet-Enanced Latent Semantic Analysis を用いているのに対し，本研究では記事のタイトルと本文全文を文書の対象とし LDA を用いている点で異なる．芹澤ら [10] は，ニュース記事の時系列データを対照に LDA を用いてトピック追跡を行っている．LDA では事前にトピック数を指定しなければならないが，トピックの類似度から適切なトピック数を推定しており，本研究もこの手法を参考にしている．しかしながら，時系列ニュースを対象にしているのに対し、本研究では対象としている文書は主題となる語をタイトルに含む全ての記事である点で異なる．

3. 提案手法

本研究では，閲覧記事の観点に着目し，記事の観点毎に対立記事を抽出する手法を提案する．以下と図 1 に提案手法の概要を示す．

(1) 閲覧記事から主題語を抽出する．
(2) 閲覧記事から記事の明示的観点，暗黙的観点を抽出する．
(3) 主題語に対してライバルとなる対立語を抽出する．
(4) (2) で抽出した記事の観点と (3) で抽出した対立語との共起度を求め，ある記事の観点に対して最も共起度の高い対立語をその記事の観点とのペアとする．
(5) 記事の観点と対立語のペアをクエリとした全ての記事を取得する．取得した全ての記事を対立記事の候補群とする．
(6) それぞれの対立記事の候補群を文書集合としてクラスタリングを行い，クエリで用いた記事の観点を持ち，且つその観点が最も値の大きいトピックに分類される記事を全て抽出する．
(7) 記事の中で主題語がクエリで用いた対立語である記事で，最も新しい記事を対立記事として抽出する．

本研究では閲覧記事および主題語をクエリとして取得する記事のニュースサイトの対象は産経ニュース [*1] とする．

3.1 主題語の抽出

本節では，ニュース記事から主題語の候補となる語 (主題語候補) を抽出し，主題語候補に重みを付与することによって主題語を抽出する手法について述べる．

3.1.1 主題語候補の抽出

ニュースの主題語は，そのニュースに頻出している単語かつ固有名詞である場合が多い [3], [4] と考え，ニュース記事の固有名詞を主題語候補として抽出する．ここで，ニュースの主題語が人名である場合，タイトルでは『苗字』＋『敬称』，本文 1 段落目では氏名である『苗字』＋『名前』＋『敬称』，本文 2 段落目以降ではタイトルと同様に『苗字』＋『敬称』で表現される場合が多い．また，長い名称の固有名詞は，本文 1 段落目は正式名称で表現されるがタイトルや本文 2 段落目以降では略称が用いられる場合が多い．また，形態素解析器では『敬称』や「第 N 回芥川賞」のように『数字』，数字の後に用いられる『助数詞』などの接尾語は名詞として扱われている場合が多い．この時，人名において『敬称』を含めた表現や長い名称において「第 N 回」まで加えた表現では主題語の意味を限定してしまう可能性がある．そこで，本文 1 段落目においては，

[*1] 産経ニュース．http://www.sankei.com/

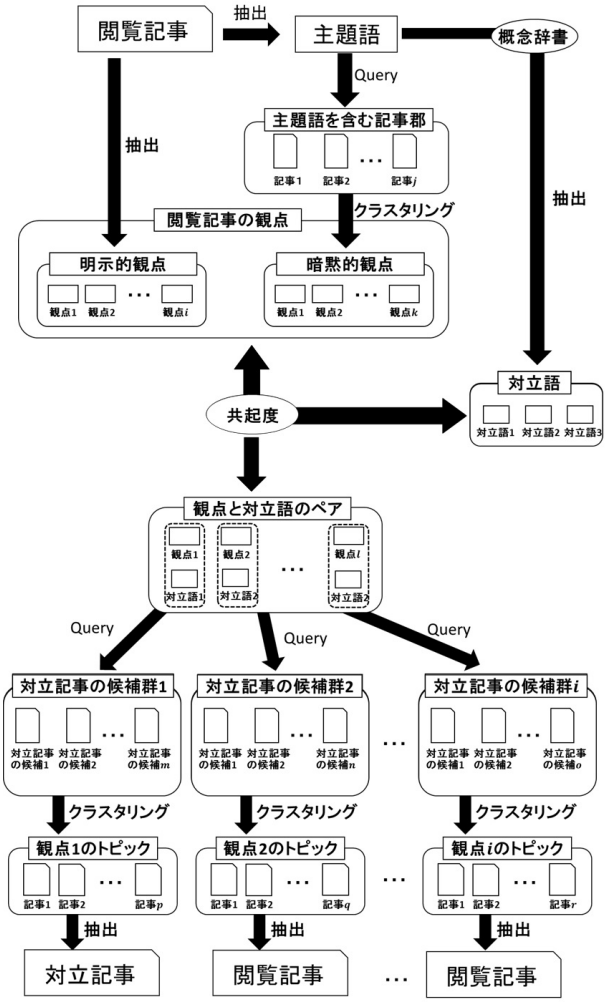

図 1　システムの概要

固有名詞の前後で出現している接尾語以外の名詞全てを結合した語を主題語候補として抽出する．

ニュースから固有名詞を抽出した後，そのニュースの主題語を決定する．ここで，一般にニュース記事は，タイトルでは記事として最も重要な情報，本文の 1 段落目ではタイトルの内容をさらに詳しい情報，そして本文の 2 段落目以降では記事の内容に付随する情報や背景などの詳細な情報を伝えている [2]．そこで本研究では記事の内容はタイトル，本文 1 段落目，2 段落目以降の順で重要であると考え，ニュース記事内に出現する固有名詞に対して出現位置を考慮した重みを付与する．

主題語候補 i に対する重み S_i を式 (1) に示す．この重み S_i の値が最も高い主題語候補を主題語と決定し抽出する．

$$S_i = \alpha \times tf_{il} + \beta \times tf_{im} + \gamma \times tf_{in} \quad (1)$$

ここで，タイトルの重みを α とし，タイトル位置 l におけるある主題語候補 i の単語の出現頻度を tf_{il} とする．同様に本文 l 段落目の重みを β，本文 l 段落目位置 m におけるある主題語候補 i の出現頻度を tf_im，本文 2 段落目以降の重みを γ，本文 2 段落目以降位置 n におけるある主題語候補 i の出現頻度を tf_{in} とする．出現位置の重み α，β，γ はそれぞれ 0.5，0.3，0.2 とする．

3.2　記事の観点の抽出

ニュースの観点には，ニュース記事に記載されている明示的観点とニュース記事には記載されていないがユーザが潜在的に知っている暗黙的観点がある．そこで，本研究ではこれら両方の観点を抽出し，記事の観点とする．

3.2.1　明示的観点の抽出

本研究で提案する明示的観点はニュース記事の特徴となる単語であることが好ましい．そこで，明示的観点となる単語もニュース記事内では主題語と同じ程度に重要であると考え，式 (1) を用いて固有名詞以外の名詞の重みを求め，閾値以上のすべての名詞を記事の観点とする．閾値は 0.6 とする．

3.2.2　暗黙的観点の抽出

暗黙的観点はそのニュース記事には記載されていないが，そのニュースの主題に関してユーザが知っているトピックであると考える，そこで，主題に対して過去のニュースからトピックを抽出し，そのトピックを暗黙的観点とする．まず，主題語をクエリとして，過去のニュース記事群を取得する．そのニュース記事群の潜在的なトピックを抽出する為に，潜在的トピック配分法（LDA: Latent Dirichlet Akkication）を用いる．LDA によって抽出されたトピックのうち，クラスタがある程度大きく，クラスタ内が疎でないクラスタのトピックを暗黙的観点とする．

3.3　対立語の抽出

本研究では，対立語は主題語と対照的な関係にある語と定義する．例えば，「野球」には「サッカー」，「自民党」には「民主党」といった語が対立語となる．2 つの語の関係に着目すると「野球」と「サッカー」は共に球技，「自民党」と「民主党」は共に日本の政党のように，対立関係にある語は共通の上位概念を持っていることが分かる．ここで，本論文では主題語と共通の上位概念を持つ語を兄弟語と定義する．また，「野球」の上位概念である球技に関して，同じ球技を上位概念にもつ兄弟語は「サッカー」だけでなく「フットサル」も挙げられる．しかしながら，「サッカー」と「フットサル」では競技人口に大きな差があり，「野球」と同程度に認知されている「サッカー」のほうが対立語として適していると考えられる．そこで本研究では対立語を，主題語と共通の上位概念を持ち，さらに同程度の認知度を持つ語と定義し，対立語を抽出する．

3.3.1　兄弟語の抽出

主題語の上位概念を取得するために概念辞書を用いる．概念辞書は様々な種類があるが新語にも対応するために，本研究では Wikipedia のカテゴリ構造からなる概念構造を概念辞書として用いる．例えば，野球では「団体競技」や「番組」など 10 語の上位概念を取得できる．これらの上位

概念のうち 1 つでも共通の上位概念を持つ語は全て主題語との兄弟語として抽出する．

兄弟語に関して，共通の上位概念を多く持つ兄弟語のほうが主題語との関係が強いと考えられる．また，どの上位概念が主題語と共通であるのかを考慮する必要もある．例えば，野球の上位概念である「団体競技」は 54 語の下位概念が存在し，「番組」は 5689 語の下位概念が存在する．この時，主題語と兄弟語において，「団体競技」における 54 語中の 2 語という関係と「番組」における 5689 語中の 2 語という関係では 54 語中の 2 語のほうが関係が強いと考えられる．このように，少数の下位概念を持つ上位概念のほうが多数の下位概念を持つ上位概念より重要である．そこで，主題語の上位概念にその下位概念の数を考慮した重みを付与し，取得した兄弟語と主題語の共通する上位概念の重みの総和を兄弟語の重みとして求める．式 (2) に主題語 s のある上位概念 U_s の重み $Sta(U_s)$ を示す．

$$Sta(U_s) = \frac{\log n}{N_s} \tag{2}$$

ここで，n は上位概念 U_s が持つ下位概念数，N_s は主題語 s の兄弟概念の数を表している．

野球における上位概念とその上位概念の下位概念数，さらに式 (2) を用いた上位概念の重みの結果を**表 1** に示す．この上位概念の重みが上位 10 件の上位概念に対して全ての下位概念を主題語の兄弟語とする．

表 1 野球の上位概念とその重み

上位概念 U_s	兄弟概念数 n	重み $Sta(U_s)$
団体競技	54	10.9
コメントするスポーツの種類	6	13.0
学生スポーツ競技・団体	1	14.9
契約選手	66	10.7
証明担当番組	235	9.4
日本の大学スポーツ競技・団体	169	9.8
番組	5689	6.2
本拠地を置くチーム・団体	7	12.9
魔球が登場する作品	39	11.2
野球を扱った作品	308	9.2

3.3.2 対立語の候補の抽出

上位概念の重みによって抽出した兄弟語から対立語の候補となる語を抽出する．式 (3) を用いて主題語 s と兄弟語 b に共通する全ての上位概念 U_s の重み $StaU_s$ の総和を兄弟語 b の重み $Rel(b)$ を求める．

$$Rel(b) = \sum_{i=0}^{n} StaU_s \tag{3}$$

ここで n は主題語 s と兄弟語 b の共通する上位概念 U_s の数を表している．

式 (3) を用いた野球の兄弟語の重みの上位 10 件の結果を**表 2** に示す．この兄弟語の重み上位 10 件を対立語の候補として抽出する．

表 2 野球の兄弟語とその重み

兄弟語 b	兄弟語の重み $Rel(b)$
バレーボール	78.0
サッカー	78.0
バスケットボール	55.2
ソフトボール	44.8
アイスホッケー	40.0
テニス	34.0
フットサル	33.6
フィールドホッケー	31.6
ゴルフ	30.4
ラクロス	26.9

3.3.3 対立語の抽出

対立語の候補となる語から対立語を抽出するため，対立語の候補の認知度を求める．我々の提案する認知度は検索結果数を用いて，主題語との検索結果数が近いほど認知度は近いと考え，式 (4) を用いて主題語 s の検索結果数 $Cogs$ と対立語の候補 c の検索結果数 $Cogc$ の認知度の対比率 $Cons, c$ を求める．

$$Con(s,c) = 1 - \frac{\{|Cog(s) - Cog(c)|\}}{max\{Cog(s), Cog(c)\}} \tag{4}$$

野球の対立語の候補 10 件の認知度の対比率を**表 3** に示す．野球に対してサッカーが取れていることが分かる．本研究では，主題語と対比率の近い上位 3 件を対立語として抽出する．

表 3 野球の対立語の候補とその対比率

対立語の候補 c	検索結果数 $Con(c)$	対比率 $Con(s,c)$
バレーボール	17700000	0.27
サッカー	57900000	0.87
バスケットボール	32700000	0.49
ソフトボール	1790000	0.02
アイスホッケー	935000	0.01
テニス	40400000	0.61
フットサル	36200000	0.54
フィールドホッケー	221000	0.00
ゴルフ	319000000	0.21
ラクロス	648000	0.01

3.4 記事の観点と対立語のペア決定

対立記事を抽出するために，対立語を含み且つ同じ観点を持つ記事の抽出を行う．ここでは，明示的観点，暗黙的観点両方を観点として用いる．この時，1 つの観点には複数の対立語が関連付けられる．そのため，観点毎に適した対立語のペアを決定する．そこで我々は，観点と対立語が頻繁に共起している場合，2 つの語の関係が深いと考え，ある観点に対して共起度の最も高い対立語をペアにする．この時，共起度を求めるために観点と対立語の共起する数が必要であるが，一般の検索結果数では共起が存在しても

ニュースとして存在しなければ対立記事を抽出できない．そこで，ニュース記事の数を用いることで最も適した共起度を求めることが出来ると考え，観点 a を含むニュース記事の数 $Cog(a)$ と対立語 r を含むニュース記事の数 $Cog(r)$，観点 a と対立語 r が共起するニュース記事の数 $Cog(a,r)$ を式 (5) の Dice 係数を用いて共起度を求める．

$$Dice(a,r) = \frac{2 \cdot Cog(a,r)}{Cog(a) + Cog(r)} \quad (5)$$

ここで，ニュース記事の数の取得には，産経ニュースのドメインを指定して得られる検索結果数を用いている．また，観点毎に対立語とのペアを決定するため，観点の数と同じ数だけペアが生成される．

3.5 対立記事の抽出

対立記事を抽出するために，対立記事の候補群を取得する．この時，対立記事は閲覧記事と同じ観点を潜在的に持っていると考えられる．そこで本節では対立記事の候補群から潜在的な観点を抽出する手法を述べる．

3.5.1 対立記事の候補群の取得

記事の観点と対立語のペアをクエリとして得られるニュース記事を対立記事の候補群として取得する．この時，検索結果として得られる記事の本文に対立語が出現せず，関連記事やページ内のランキングに対立語が出現しているため取得される場合がある．このような場合，検索結果のスニペットにはクエリとなる対立語が出現しない．そこで，スニペットに対立語が出現するニュース記事を対立記事の候補群として最大 300 件取得する．

同様にして全てのペアに対して対立記事の候補群を取得する．

3.5.2 対立記事の決定

閲覧記事の観点毎に対立記事の候補群から対立記事を決定する．しかしながら対立記事は内容が閲覧記事と類似しているとは限らないため，単純に記事の内容で類似度計算を行っては対立記事となる記事の類似度が高いとは限らない．そこで，我々は記事の観点に着目し，閲覧記事と観点が類似している記事が対立記事となると考え，対立記事の候補群それぞれの記事から観点を抽出し，閲覧記事の観点との類似度を求め，最も類似度の高い記事を対立記事とし，提示する．この時，類似度計算には *cos* 類似度を用いる．

4. 評価実験

主題語と記事の明示的観点の抽出の有用性を計る為に，評価実験を行った．実験データは産経ニュースからランダムに選んだ記事 20 件とし，被験者は 20 代男女 8 名とする．

被験者はシステムを使用して，主題語の抽出結果と記事の明示的観点の抽出結果に対してそれぞれ 5 段階評価を行った．評価値は，抽出された結果が最も適していれば評価値を 5，最も適していなければ 1 とした．

主題語に関する評価値の割合を図 2 に示す．図 2 より評価値 4〜5 の割合が 95%となり，閲覧記事から主題語が適切に抽出できていると言える．主題語の評価が低かった例としては「ＢＳＥ発生、ノルウェー産牛肉輸入停止　厚労省発表」に関する記事で，主題語は「ノルウェー」が抽出された．評価の低かった原因としては，適切な主題語が「ノルウェー」だけでなく「BSE」や「牛肉」，「厚生労働省」など複数考えられる点や，そのうち「牛肉」という語が一般名詞である点が挙げられる．そのため，主題語を抽出する際は一般名詞も考慮する必要があると考えられる．

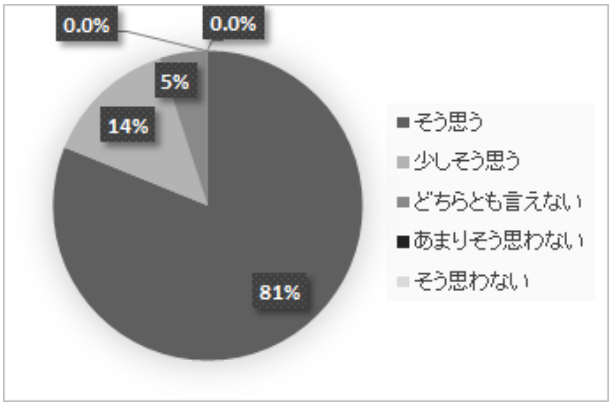

図 2 主題語に関する評価値の割合

閲覧記事の記事アスペクトに関する評価値の割合を図 3 に示す．図 3 より評価値 4〜5 の割合が 52%となった．記事アスペクトの評価の高かった例としては「鹿児島県知事、川内原発再稼働に同意を表明」に関する記事で，記事アスペクトが「稼働」や「同意」，「知事」，「原発」となり，記事アスペクトのみで記事の概要を把握できるような語が抽出されたためだと考えられる．評価の低かった例としては，「ＫＤＤＩが春モデル発表　ボルテ対応スマホを 5 機種投入　アンドロイド搭載のガラケーも」に関する記事で，記事アスペクトが「機種」や「対応」，「発売」，「モデル」，「向け」，「搭載」となり，特に「対応」，「発売」，「向け」といった語の評価が低かった．原因として，評価の低かった記事

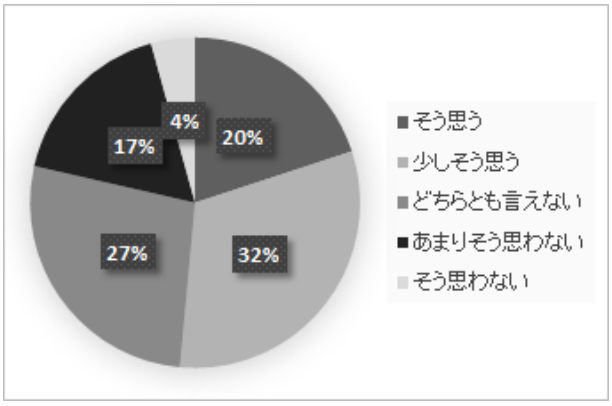

図 3 明示的観点に関する評価値の割合

アスペクトはこの記事で抽出された主題語である「スマホ」特有の語ではなく，様々な語に対して用いられやすい語，つまり汎用性の高い語であることが分かった．そのため，汎用性の高い語には重みを小さくすることや，前後の語と組み合わせることで「向け」ではなく「子供向け」など汎用性を低くする必要があると考えられる．

5. まとめと今後の課題

ニュース記事の重要性の理解を支援することを目的に，ユーザの閲覧している記事に対する観点毎の対立記事を抽出する手法を提案した．具体的には，まず閲覧記事から主題語を抽出し，ニュースの観点として，ニュース記事に記載されている明示的観点とユーザがすでに知っていてニュースに記載されていない暗黙的観点を抽出し，主題語と観点から対立語を抽出する．さらに閲覧記事の観点と対立語のペアを決定し，対立記事を抽出した．

今後の課題として，対立記事抽出の評価実験を行う．また主題語と閲覧記事の観点抽出で用いた重みをさまざまな値を使って抽出し，適合率を計る実験を行うと共に主題語と観点の評価実験を行う．また，ニュース記事を取得する対象としているサイト数を増やし，対立記事抽出の制度の向上と閲覧記事の多様性を図る．さらに，対立語を抽出する際，主題語と共通する上位概念に対して記事の内容を考慮することで，より主題語と関係の強い対立語の抽出を行う． そして，ユーザインタフェースの作成を行う．

謝辞 本研究の一部は JSPS 科研費 26330347 及び，私学助成金 (大学間連携研究補助金) の助成によるものです．ここに記して謝意を表します．

参考文献

[1] 池田大介, 藤木稔明, 奥村学. ”blog とニュース記事の自動対応付け”. 言語処理学会第 11 回年次大会論文集, pp.1030-1033, 2005.
[2] 北山大輔, 角谷和俊. ” ニュースアーカイブのためのコンテンツ構成順序を用いた比較ニュース検索”. 日本データベース学会 letters, Vol. 6, No. 1, pp. 169-172, jun 2007.
[3] Keisuke Kiritoshi and Qiang Ma. ”Named entity oriented related news ranking”. In Database and Expert Systems Applications, pp. 82-96. Springer, 2014.
[4] 田中祥太郎, ヤトフトアダム, 田中克己. ” ニュース記事の理解支援のための背景知識抽出と補完”. 研究報告データベースシステム（DBS）, Vol. 2014, No. 17, pp. 1-6, jul 2014.
[5] 佐藤吉秀, 川島晴美, 佐々木努, 奥雅博. ” 時系列ニュース記事における最新話題語抽出方法”. 情報処理学会研究報告. NL, 自然言語処理研究会報告, Vol. 168, pp. 1-6, jul 2005.
[6] 菊地匡晃, 岡本昌之, 山崎智弘. ” 階層型クラスタリングを用いた時系列テキスト集合からの話題推移抽出”. 日本データベース学会論文誌, Vol. 7, No. 1, pp. 85-90, 2008.
[7] 高橋佑介, 横本大輔, 宇津呂武仁, 吉岡真治. ” ニュースにおけるトピックのバースト特性の分析”. 研究報告自然言語処理（NL）, Vol. 2011, No. 6, pp. 1-6, nov 2011.
[8] David M Blei, Andrew Y Ng, and Michael I Jordan. ”Latent dirichlet allocation”. the Journal of machine Learning research, Vol. 3, pp. 993-1022, 2003.
[9] 吉田稔, 中川裕志, 石田智也. ” ニュース記事クラスタリングによる取引高予測の試み”. 人工知能学会全国大会論文集, Vol. 25, pp. 1-4, 2011.
[10] 芹澤翠, 小林一郎. ” 潜在的ディリクレ配分法に基づくトピック類似度を考慮したトピック追跡”. 第 4 回 DEIM フォーラム論文集, 2011.