甲南大学大学院
自然科学研究科
知能情報学専攻
修士論文 No. 183

観点に基づくWebニュース対立記事抽出手法の提案

Exctracting Rival-news from Web News Articles based on Aspect of Subject Term of the News

2017年3月
大原 正章

甲南大学大学院 自然科学研究科

要旨

近年，インターネットの普及により様々なニュースサイトが利用されている．人々はこのようなニュースサイトからいつでもリアルタイムにニュースを取得することが可能である．そのため，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大性を把握できない場合がある．このような場合，閲覧記事に対して対立関係にあるニュース記事と比較することが出来れば，閲覧記事の重要性を知ることが容易になると考えられる．そこで本研究では Web ニュースの理解支援の一つとして，閲覧記事から対立する記事を抽出する手法の提案を行う．具体的には，Web ニュースの主題となる語を抽出し，その主題に対してニュースに記載されている明示的観点と主題の背景となる暗示的観点を抽出する．そして各々の観点毎にユーザの閲覧している記事との対立語を抽出し，その対立語を用いて対立記事を抽出する手法を提案する．

Summary

Nowadays, there are many kinds of news sites on the Internet. People can get news articles from the news sites in real time. Then using Web news is beneficial for the users to get latest news articles. However, when people browse web news, they sometimes can not understand how important of the news articles. In this way, we consider that if we can compare the news article with rival news article, we can know importance of the news article easily. Then we propose the method which extracts rival news from user's browsing news automatically. In particular, we first extract subject term from the Web news. We next extract two kinds of aspect which is explicit aspect and implicit aspect. Then we extract rival news articles by using explicits and implicit aspects.

目 次

図 目 次

表 目 次

1　はじめに

近年，インターネットの普及により様々なニュースサイトが利用されている．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可能なため，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 200 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作家が書いた本の発行部数を知らない場合，この 200 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，多くのユーザが 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と考えられる．しかしながら，関連記事の多くは閲覧している記事の内容に関する過去に報道された記事である場合や，ユーザが閲覧している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，関連記事を閲覧しても，元のニュース記事の重要性を理解することは困難である場合が多数ある．

このような場合，ユーザが閲覧している記事に対して過去に別の芸人が書いた書籍の発行部数に関するニュース記事や，別の芥川賞作家が書いた書籍の発行部数に関するニュース記事と比較することができれば，ユーザの閲覧している記事の重要性を理解する手助けになると考えられる．この時,「又吉」に対して比較対象となる芸人や芥川賞作家はライバルとなり，そのライバルの書いた本の発行部数に関するニュース記事はユーザの閲覧している記事と対立関係にあると考えられる．

そこで，ユーザの閲覧している記事と対立関係にあるニュース記事を提示することにより，その閲覧している記事の重要性を理解することが可能であると考え，このユーザの閲覧している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．

本研究では，ユーザの閲覧している記事を閲覧記事と呼び，この閲覧記事と対立関係にあるニュース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人物や組織を表す語である主題が存在する．本研究では，記事内の主題となる語を主題語と呼ぶ．また，閲覧記事と対立記事は対立関係にあると考えられるが，この時，主題語と対立記事の主題も対立関係にあると考えられる．このような対立記事の主題であり主題語と対立関係にある語を対立語と呼ぶ．

ここで，主題語は複数の観点を持つ場合がある．例えば，上記のニュース記事の例では「又吉」が主題語となり，この記事における「又吉」という語は「芸人」や「芥川賞」等の観点を持つと考えられる.「芥川賞」という語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「又吉が芸人である」という情報は人々が知識として潜在的に知っている観点であるため,「芸人」という語は記事中に記載されているとは限らない．このような場合，暗黙的観点は閲覧記事に記載されていない場合でも他の主題語に関する記事では記載される場合が多いと考えられる．そこで本研究では，記事中に記載されている観点を明示的観点，閲覧記事以外の主題語を含む記事に多く記載されている観点を暗黙的観点と呼ぶ．この時,「芥川賞」を観点とすると対立語は他の芥川賞作家であり，対立記事は「他の芥川賞作家の本の発行部数」に関する記事になる．また,「芸人」を観点とすると対立語は他の芸人であり，対立記事は「他の芸人の本の発行部数」に関する記事となる．このように観点によって対立語と対立記事は異なる．

そこで本研究では，閲覧記事から主題語及び明示的観点と暗黙的観点を抽出し，これらを用

いて観点毎に対立語と対立記事を抽出する手法について提案する．

具体的には，まず閲覧記事から主題語と明示的観点を抽出する．次に主題語を用いて暗黙的観点を抽出する．ここで，明示的観点に基づく対立記事を抽出するため，主題語から対立語の抽出を行い，対立語と明示的観点を用いて対立記事の抽出を行う．また，暗黙的観点に基づく対立記事を抽出するため，暗黙的観点を用いて対立語の抽出を行い，対立語と閲覧記事に出現する名詞を用いて対立記事の抽出を行う．

以下，第 2 章では関連研究について，第 3 章では対立記事の抽出手法の概要，第 4 章で主題語の抽出，第 5 章で観点の抽出，第 6 章で対立語の抽出，第 7 章で対立記事の抽出，第 8 章で実験と考察，そして第 9 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だけでは得られない情報を提示する手法を提案している．この手法ではニュース記事や blog をベクトル変換することでその類似度から対応付けを行う．本研究でもニュース記事同士をベクトル変換し，類似度の高い記事を対応付けを行っているが，その対象がニュース記事同士である点で異なる．北山ら [2] 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．馬 [4] らの研究では，情報補完の観点から，あるコンテンツに対する異なった視点のコンテンツや，より詳細に説明がなされているコンテンツの検索手法を提案している．本研究でも，あるコンテンツに対する異なった視点に着目しているが，その視点を用いてコンテンツのより詳細な情報を抽出せず観点を軸として対立語を抽出している点で異なる．小林ら [6] は，ニュース記事から専門用語を抽出し，その専門用語に対する解説を，ソーシャルビデオ情報，ツイート情報，関連プロダクト情報といった多面的な関連情報マルチコンテンツ) から取得し，それらをユーザが利用しやすいインタフェース上に提示する手法を提案している．本研究でも主題語に関する多面的な情報を観点として抽出している点で類似しているが，抽出する対象をニュースメディアのみである点で異なる．田中ら [5] は，あるニュース記事の背景知識を抽出し提示する手法を提案している．この手法で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，それらエンティティ間の関係を示すことで背景知識の補完を目的としている．本研究でもニュース記事内に出現する人物や組織等の固有名詞に着目しているが，それら固有名詞から主題語を抽出し，一般名詞からは観点を抽出している点で異なる．

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 [7] 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 [8] は，ニュース記事などの時系列テキスト集合において階層型クラスタリングを適用し話題語を抽出している．この時，抽出されるクラスタから話題語を抽出するため，C-value 法を用いてキーワード郡を抽出し，そのクラスタに分類される文書集合から求めた idf 値より話題語の抽出を行っている．高橋ら [9] 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 Allocation）[10] 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 [11] 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 [12] 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　提案手法の概要

本研究では，ユーザの閲覧している記事の内容に関する理解支援を目的とし，ニュースの主題とその観点に着目し，閲覧記事と対立関係にある記事を自動で抽出して提示する手法の提案を行う．本研究ではこのように閲覧記事と対立関係にある記事を対立記事と呼び，抽出を行う．

3.1　主題語の定義

ニュース記事には記事内で話題の中心となる人物や団体，地名などを指す固有名詞が存在する場合が多い．このような語を主題語と呼ぶ．本研究では主題語を以下の定義にしたがって抽出する．

- ニュース記事内で重要な語である．
- 固有名詞である．

3.2　対立語の定義

例えば「又吉直樹が芥川賞を受賞した書籍の発行部数が 200 万部を突破した」という閲覧記事に対して,「又吉直樹」は同じ芥川賞を受賞した「羽田圭介」や，同じ芸人である麒麟の「田村裕」と比較することができると考えられる．本研究ではこのような主題語と比較できると考えられる語を対立語と呼ぶ.

この時,「又吉直樹」のに対する「羽田圭介」や「田村裕」の関係に着目すると,「又吉直樹」と「羽田圭介」は同じ「芥川賞作家」であり,「又吉直樹」と「田村裕」は同じ「お笑い芸人」で

ある．このように対立関係にある語同士は共通の上位概念を持つと考えられる． さらに上位概念の「お笑い芸人」に着目すると，共通の上位概念として持つ語は「田村裕」だけでなく「明石家さんま」「ビートたけし」などが挙げられる．しかしながら,「又吉直樹」に対して「明石家さんま」や「ビートたけし」では知名度に大きな差があるため，対立語として適切ではないと考えられる．そこで対立語の定義を以下に示す．

- 主題語と共通の上位概念を持つ．
- 主題語と同程度の認知度を持つ．

3.3 観点の定義

例えば,「又吉直樹」の対立語は「羽田圭介」や「田村裕」が挙げられる．この時，対立語は「芥川賞」や「芸人」など，観点によって対立語は異なる場合がある．さらに，観点によって対立語が異なる場合は対立記事も異なると考えられる．

この時,「芥川賞」という語は閲覧記事内に出現する重要な名詞であると考えられる．本研究ではこのような語を明示的観点と呼び，抽出する．また,「又吉直樹は芸人である」という情報は誰もが知っている情報であると考えられる．このような場合，閲覧記事内に「芸人」という語が出現しない場合があると考えられるが,「芸人」という語は「又吉直樹」と強い関係を持つと考えられる．本研究ではこのような語を暗黙的観点と呼び抽出を行う．以下に明示的観点の定義を示す．

- ニュース記事内で重要な語である．
- 一般名詞である．

また，以下に暗黙的観点の定義を示す．

- 閲覧記事の主題語と関係の強い語である．
- 一般名詞である．

3.4 提案手法の手順

本研究では閲覧記事から観点毎に異なる対立記事の抽出を行う．以下と図 1 に提案手法の手順を示す．

(1) 閲覧記事から主題語を抽出する．
(2) 閲覧記事から，明示的観点を抽出する．
(3) (1) で抽出した閲覧記事の主題語を用いて暗黙的観点を抽出する．
(4) それぞれの明示的観点を用いて対立語と対立記事を抽出する．
(5) それぞれの暗黙的観点を用いて対立語と対立記事を抽出する．
(6) (4) と (5) で抽出した対立記事を提示する．

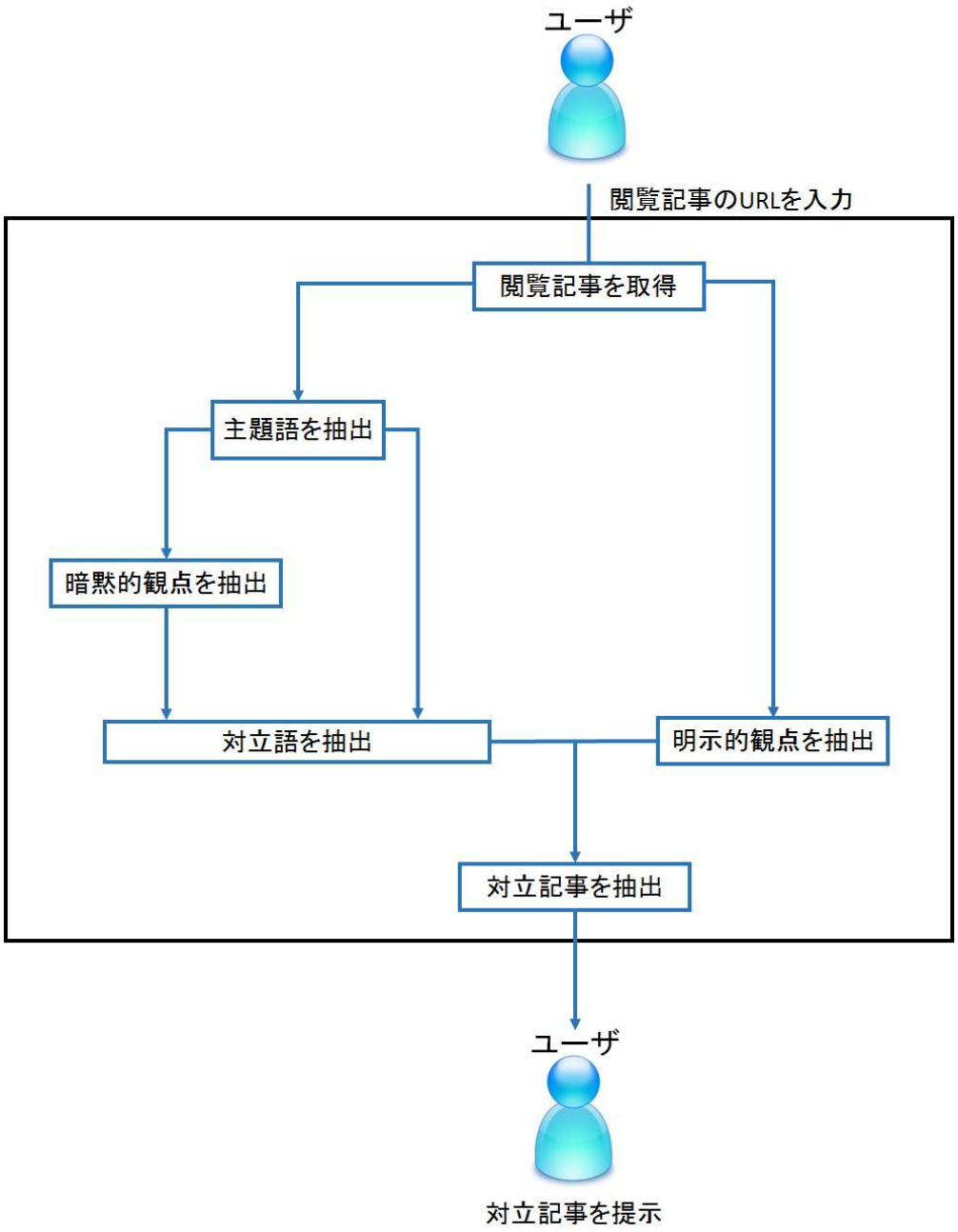

図 1: 提案手法のシステムフロー

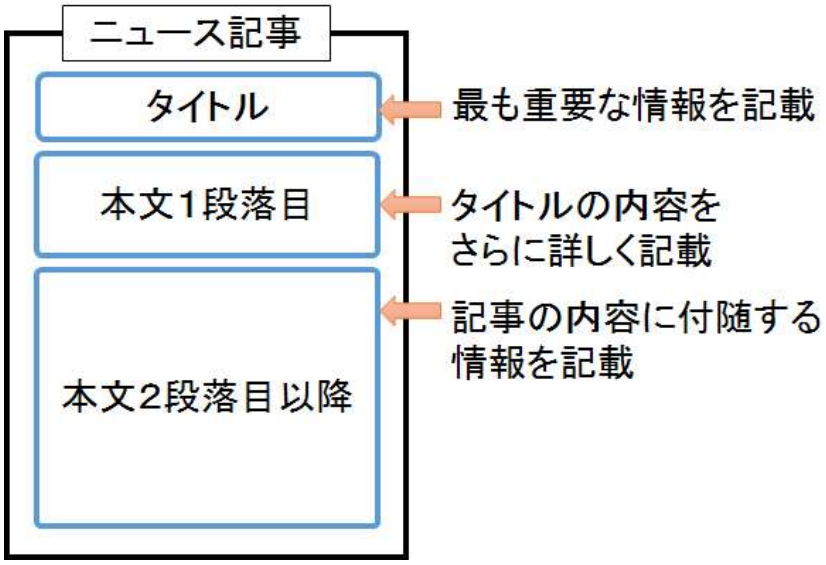

図 2: ニュース記事の構成

4 主題語の抽出

ニュース記事には記事内で話題の中心となる人物や団体，地名などを指す固有名詞が存在する場合が多い．このような記事の話題の中心となる固有名詞を本研究では主題語と呼び抽出する．

ここで，一般にニュース記事は図 2 に示すように，タイトルはその記事を顕著に表す語で構成されており，本文の 1 段落目ではタイトルの内容についてさらに詳しい情報が記載されている．さらに本文 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が記載されている [2]．そこで本研究ではこれら記事の構造に着目し，ニュース記事のタイトルと本文に出現する人名や団体，地名の固有名詞 i に対して式 (1) に示すように，出現位置と出現頻度を考慮して単語毎の重み S_i の値を求める．そしてこの S_i が最も高い単語を主題語とする．この時，タイトルの重みを α とし，タイトル位置 l のある固有名詞 i の単語の出現頻度を $tf_{i,l}$ とする．同様に本文 1 段落目の重みを β，本文 1 段落目位置 m のある固有名詞 i の出現頻度を $tf_{i,m}$，本文 2 段落目以降の重みを γ，本文 2 段落目以降位置 n のある固有名詞 i の出現頻度を $tf_{i,n}$ とする．

$$S_i = \alpha tf_{i,l} + \beta tf_{i,m} + \gamma tf_{i,n} \tag{1}$$

この時，タイトルに出現する単語は略称で記載されている場合がある．例えば「安倍晋三」という人物に対してタイトルでは「安倍総理」もしくは「安倍首相」と表現される場合が多く，形態素に分割した際に得られる「安倍」という語だけでは「安倍晋三」と判断することは困難である．このような場合，記事内の本文で最初に「安倍晋三」について言及している文章には「安倍晋三内閣総理大臣」など正式名称で記載されており，以降の表現はタイトルに沿った表記がされている場合が多い．このように人物名や会社名などの固有名詞がタイトルで簡略化された表現がされている場合でも，本文 1 段落目では正式名称が記載される．そこで略称で記載さ

れている単語と正式名称の単語を同一の単語と見なすために，タイトルや本文 2 段落目以降に出現する単語に対し，その単語を含みかつその単語より文字数が多い語が本文 1 段落目に存在すれば，本文 1 段落目の単語が正式名称であるとして置き換える．この時，人名に限り「さん」や「様」などの敬称,「総理」や「代表取締役」などの役職名は除去することで人物名のみの抽出を行う．

5 観点の抽出

ニュース記事における主題語の観点には，ニュース記事に記載されている語である明示的観点と，ニュース記事には記載されていない場合もあるがユーザが潜在的に知っている語であ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，主題語の観点とする．

5.1 明示的観点の抽出

例えば,「又吉が芥川賞を受賞した書籍の発行部数が 200 万部を突破した」という閲覧記事に対して,「又吉」と同じ芥川賞を受賞した人物のニュース記事と比較することができれば，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．この時，この「芥川賞」という語は閲覧記事内で記載されている語であり，閲覧記事内において重要な語であると考えられる．このような語を本研究では明示的観点と呼び，抽出を行う．

具体的には，閲覧記事内に出現する一般名詞を対象に式 (1) を用いて語の値を算出し，その値が閾値以上の単語をすべて明示的観点として抽出する．

ここで，式 (1) の重み α，β，γ の値を決定するため,「トランプ氏国家情報長官にコーツ前上院議員を起用」,「韓国北朝鮮のＩＣＢＭ発射はいつでもありうる」,「政府少女像設置への対抗措置駐韓国大使あす一時帰国へ」の 3 件のニュースを用いて α，β，γ に対して α+β+γ=1.0 となる 3 つのパターンで得られる値 S_i を表 1〜3 に示す．

表 1: 名詞の重み計算の結果 1

α=β=γ		α=0.6，β=0.3，γ = 0.1		α=0.8，β=0.15，γ=0.05	
情報	0.0387962	情報	0.046916	情報	0.043458
コーツ	0.0225925	コーツ	0.032333	コーツ	0.036166
議員	0.0202777	議員	0.03025	議員	0.035125
トランプ	0.0202777	トランプ	0.03025	トランプ	0.035125
国家	0.0202777	国家	0.03025	国家	0.035125
上院	0.0202777	上院	0.03025	上院	0.035125
国家情報	0.0179629	国家情報	0.028166	国家情報	0.034083
長官	0.0179629	長官	0.028166	長官	0.034083
情報機関	0.0138888	情報機関	0.0125	情報機関	0.00625

表 1〜3 より，重みの値 α，β，γ によって名詞の値 S_i は大きく変化しないことが分かる．これは，重要な単語はタイトルや本文 1 段落目，本文 2 段落目以降のすべての位置において出現しやすいからであると考えられる．しかしながら，これら出現位置についてはタイトルに出現

表 2: 名詞の重み計算の結果 2

α=β=γ		α=0.6，β=0.3，γ = 0.1		α=0.8，β=0.15，γ=0.05	
北朝鮮	0.044566	北朝鮮	0.054395	北朝鮮	0.060531
ICBM	0.041391	ICBM	0.053443	ICBM	0.060054
韓国	0.041391	韓国	0.053443	韓国	0.060054
発射実験	0.01916	発射実験	0.013443	発射実験	0.006721
大陸間	0.015995	大陸間	0.012490	大陸間	0.006245
韓国軍	0.015995	韓国軍	0.012490	韓国軍	0.006245

表 3: 名詞の重み計算の結果 3

α=β=γ		α=0.6，β=0.3，γ = 0.1		α=0.8，β=0.15，γ=0.05	
韓国	0.041648	韓国	0.039972	韓国	0.036652
政府	0.029509	政府	0.030925	政府	0.032129
像	0.027944	像	0.030455	像	0.031894
大使	0.026379	大使	0.029986	大使	0.031659
少女	0.026379	少女	0.029986	少女	0.031659
帰国	0.023250	帰国	0.029047	帰国	0.031190
一時	0.023250	一時	0.029047	一時	0.031190
対抗措置	0.021685	対抗措置	0.028577	対抗措置	0.030955
問題	0.019963	設置	0.02	設置	0.026666

する語が最も重要な単語であると考えられるため，重みの値 α，β，γ を全て同じ値にすることは不適切である．また，タイトルの単語の重みが極端に大きい場合，明示的観点を抽出する際にタイトルに出現する名詞がほぼ必ず抽出される可能性がある．そのため，重みの値 α，β，γ はそれぞれ 0.6，0.3，0.1 とする．また，主観による実験より，明示的観点における閾値は 0.02 とする．

表 4 で示すニュース記事の例から算出した固有名詞の値を表 5 に示す．この結果より表 4 のニュース記事の主題語は「又吉直樹」となる．

また，表 4 で示すニュース記事の例から算出した一般名詞の値を表 6 に示す．この結果より表 4 のニュース記事の明示的観点は「火花」,「コメント」,「たくさん」,「本」,「夏」,「突破」,「芥川賞」となる．

5.2 暗黙的観点の抽出

例えば,「又吉」に関するニュースに対して同じ芸人である人物のニュース記事と比較することが出来れば，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．この時,「又吉は芸人である」という情報は人々が既に知っている知識であるため，閲覧記事内に出現しない場合があると考えられる．このような語を暗黙的観点と呼ぶ．この時，暗黙的観点は主題語と関係が強い語であると考えられる．そのため，例えば「芸人」という語が閲覧記事内に出現しない場

表 4: ニュース記事の例

タイトル	又吉さん「火花」が２００万部突破 「この夏、たくさん本を読む」とコメント
本文１段落目	文芸春秋は４日、お笑いコンビ「ピース」の又吉直樹さん（３５）による第１５３回芥川賞受賞作「火花」について、さらに４０万部の増刷を決めたと発表した。累計発行部数は１８刷２０９万部となる。
本文２段落目以降	全国的な品薄状態は徐々に解消されつつある、というが、「８月下旬に芥川賞の贈呈式が行われるため大きな話題となることが予想される。その前に書店の在庫を最大化しておきたい」（同社）と理由を説明している。...

表 5: 抽出された固有名詞とその値

固有名詞 i	値 S_i
又吉直樹	0.052941
ピース	0.035294
お笑いコンビ	0.035294
文芸春秋	0.035294
マディソン郡の橋	0.005882

表 6: 抽出された一般名詞とその値

一般名詞 i	値 S_i
火花	0.0647058
コメント	0.0411764
たくさん	0.0411764
本	0.04117646
夏	0.0411764
突破	0.0352941
芥川賞	0.023529
増刷	0.0176470
同社	0.0176470
部数	0.0117647

合でも，他の「又吉」を含むニュース記事には「芸人」という語が出現しやすいと考えられる．また,「芸人」という語は「又吉」の上位概念である．そこで本研究では，以下に示すように 2 つの暗黙的観点の抽出手法を提案する．

- 主題語を含むニュース記事群からトピックを抽出することでそのトピックから暗黙的観点

を抽出する手法

- 主題語の上位概念から暗黙的観点を抽出する手法

以下にそれぞれの具体的な手法について述べる．

5.2.1 トピック抽出に基づく暗黙的観点の抽出手法

暗黙的観点は閲覧記事に存在しない場合がある．しかしながら，主題語を含む他の記事には暗黙的観点が出現しやすいと考えられる．そこで，主題語を含む記事郡から潜在的なトピックを抽出することで暗黙的観点を抽出する．

トピック抽出を用いた暗黙的観点抽出の概要を以下と図 3 に示す．

(1) 主題語とサイト指定検索を用いてニュース記事 300 件を取得する．

(2) 取得したニュース記事それぞれから名詞を抽出する．

(3) 取得した名詞を用いてトピックを抽出する．

(4) 得られたトピックから暗黙的観点を抽出する．

具体的には，まず主題語をクエリとして閲覧記事以外のニュース記事を 300 件取得する．この時，ニュース記事の Web ページのみを取得するため，ニュースサイトのドメインを用いたサイト指定検索を行う．次に取得したニュース記事郡群から各々の名詞を抽出し，その名詞を用いてトピック抽出を行う．この時，ニュース記事群の潜在的なトピックを抽出する為に，潜在的ディリクレ配分法（LDA：Latent Dirichlet Allocation）を用いる．LDA とは，1 つの文書が複数のトピックから構成されると仮定した確率モデルである．この各クラスタに含まれる記事を多く持つクラスタのトピックを暗黙的観点として抽出する．本研究では 10 件以上の記事を持つクラスタのトピックを暗黙的観点として抽出する．

以下に「又吉直樹」を用いて抽出した結果の一部を表 7 に示す．この結果より暗黙的観点は「火花」,「小説」,「記者」,「相方」,「作品」となる．

5.2.2 概念構造に基づく暗黙的観点の抽出手法

暗黙的観点となる語は主題語の上位概念であると考えられる．また，暗黙的観点は主題語と強い関係を持つことから，主題語と共起しやすい語であると考えられる．そこで，主題語の上位概念のうち，主題語との共起度が高い語を暗黙的観点として抽出する．

具体的には，まず概念辞書を用いて主題語の上位概念を取得する．概念辞書には，多くの情報量を持ち新語にも対応している Wikipedia のカテゴリー構造から生成される辞書を用いる．

この時，主題語と関係が強いとは言えない上位概念が存在する．そこで主題語と関係の強い上位概念のみ抽出するため，共起度を用いて計算を行う．ここで，共起度を求めるには，Dice 係数，Jaccard 係数，Simpson 係数など様々な尺度があるが，本研究では式 (2) に示す Dice 係数を用いる．

$$D_{su} = \frac{2 \times R_{su}}{R_s + R_u} \tag{2}$$

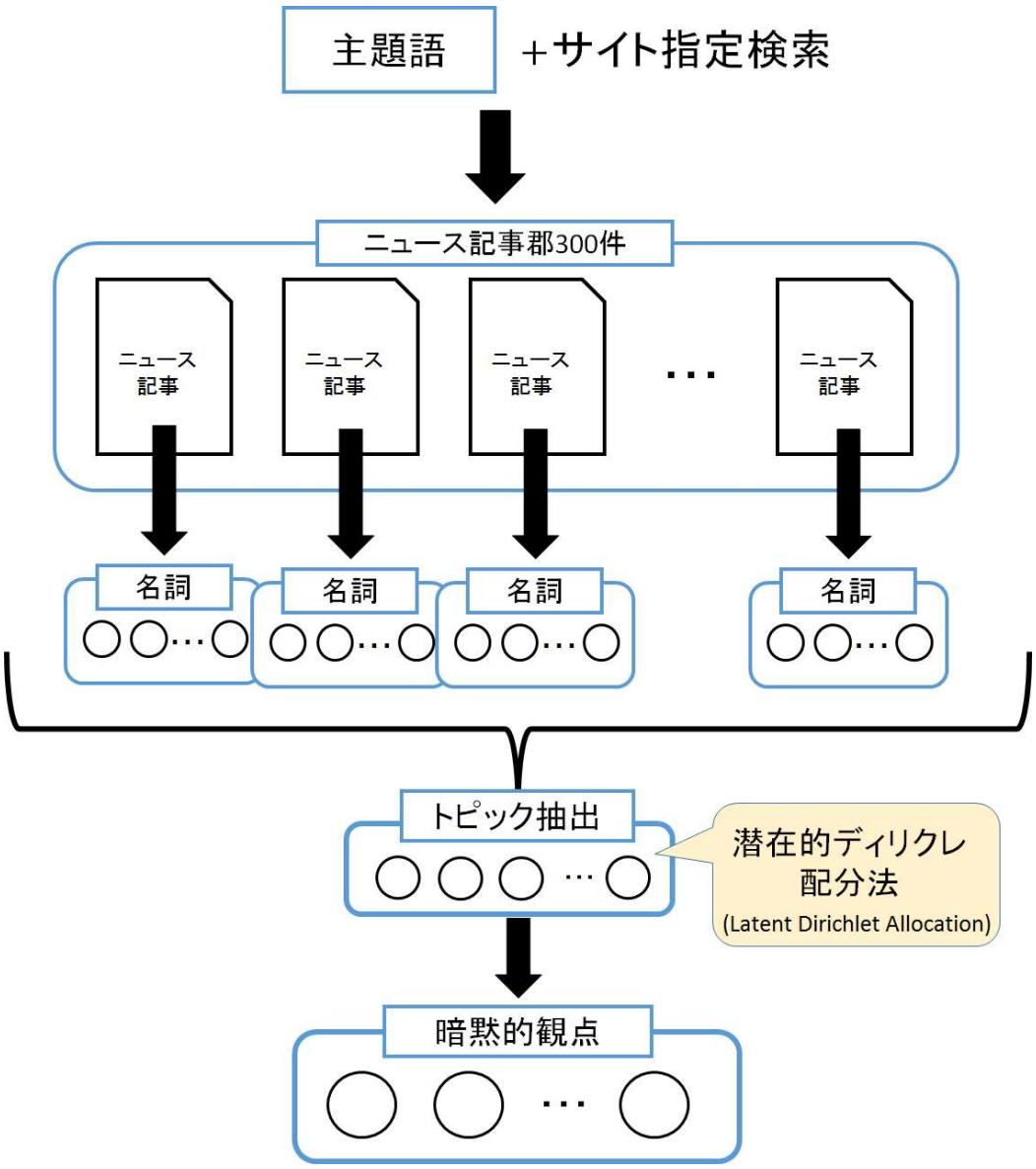

図 3: トピック抽出に基づく暗黙的観点抽出の概要

表 7: クラスタリング結果の一部

トピック	クラスタに含まれる記事数
火花	16
小説	14
記者	12
相方	10
作品	10
芸人	8
先輩	6
俳句	6
番組	6
先生	6
撮影	5
作家	4
CM	4
制作	3

ここで，主題語 s の検索結果数を R_s，主題語 s のある上位概念 u の検索結果数を R_u，主題語 s と主題語 s のある上位概念の両方を含む検索結果数を R_{su}，主題語 s と主題語 s のある上位概念の共起度を D_{su} とする．主観による実験より，この共起度の値 D_{su} が 0.02 以上の上位概念を暗黙的観点として抽出する．

主題語「又吉直樹」の検索結果数を表 8,「又吉直樹」の上位概念それぞれの検索結果数を表 9，さらに「又吉直樹」とそれぞれの上位概念を組み合わせた時の検索結果数を表 10 に示す．また，この時の共起度を表 11 に示す．この結果より「又吉直樹」の暗黙的観点は「芥川賞受賞者」「日本の小説家」「俳人」「お笑い芸人」となる．

表 8: 主題語「又吉直樹」の検索結果数

主題語	検索結果数
又吉直樹	378000

この時,「又吉直樹」とコンビ名である「ピース」は関係が強いと考えられるが，共起度が低い結果となった．原因として「ピース」が指す意味はお笑いコンビ名だけでなく写真撮影を行う際の合図であったり数を数える単位の一つであることから検索結果数が非常に大きくなり共起度が低くなったと考えられる．

6 対立語の抽出

例えば,「又吉が芥川賞を受賞した書籍の発行部数が 200 万部を突破した」という閲覧記事の対立記事は，同じ芥川賞受賞者に関する記事や同じ芸人に関する記事となり，観点毎に対立語

表 9: 主題語「又吉直樹」の上位概念の検索結果数

上位概念	検索結果数
お笑い芸人	1400000
日本のタレント	2710000
ピース	11800000
日本の小説家	1200000
芥川賞受賞者	54500
俳人	895000
吉本興業	749000
大阪府出身の人物	594000
1980 年生	1320000
存命人物	1410000

表 10: 主題語「又吉直樹」かつ上位概念の検索結果数

主題語 + 上位概念	検索結果数
又吉直樹 + お笑い芸人	25900
又吉直樹 + 日本のタレント	4640
又吉直樹 + ピース	108000
又吉直樹 + 日本の小説家	4880
又吉直樹 + 芥川賞受賞者	3130
又吉直樹 + 俳人	5950
又吉直樹 + 吉本興業	13200
又吉直樹 + 大阪府出身の人物	18
又吉直樹 + 1980 年生	38
又吉直樹 + 存命人物	41

と対立記事は異なる．そこでまず観点毎に対立語の抽出を行う．

6.1 明示的観点に基づく対立記事抽出のための対立語抽出手法

閲覧記事の主題語を用いて対立語の抽出を行う．以下に対立語抽出の手順を示す．

(1) 主題語の上位概念を取得する．

(2) 取得した上位概念から兄弟語を抽出する．

(3) 兄弟語が持つ上位概念の重み計算を行う．

(4) 兄弟語から対立語候補を抽出する．

(5) 対立語候補から認知度を用いて対立語を抽出する．

表 11: 主題語「又吉直樹」と上位概念の共起度

上位概念	共起度
芥川賞受賞者	0.165486
日本の小説家	0.137230
俳人	0.030358
お笑い芸人	0.024333
日本のタレント	0.019778
ピース	0.016267
大阪府出身の人物	0.012911
吉本興業	0.012540
1980 年生	0.004732
存命人物	0.001540

6.1.1 兄弟語の抽出

本研究では，主題語と共通する上位概念を持つ語を兄弟語と呼び，抽出を行う．

具体的には，まず Wikipedia のカテゴリー構造を利用した概念辞書を用いて主題語の上位概念を取得する．Wikipedia とは誰もが自由に編集に参加できるインターネットサイトである．そのためニュース記事のような速報性の高い物事に対しても対応している場合が多く，本研究の概念辞書に適していると考えられるため，これを用いる．この時得られる主題語の上位概念が持つ下位概念語を主題語の兄弟語として抽出する．

6.1.2 対立語候補の抽出

例えば「又吉直樹」の上位概念には「お笑い芸人」や「存命人物」などが存在する．これら上位概念すべてから兄弟語を抽出した場合,「お笑い芸人」の下位概念数は 1432 件が存在し,「存命人物」は 72774 件の下位概念語が存在するため兄弟語の数が膨大になると考えられる．そこで兄弟語からより主題語と関係の強い語を対立語の候補として抽出する．

ここで,「又吉直樹」の上位概念である「お笑い芸人」と「存命人物」を比較すると，それぞれが持つ下位概念語数に大きな差がある．このような場合，下位概念語数の多い上位概念の下位概念語同士の関係より，下位概念語の少ない上位概念の下位概念語同士の関係のほうが強いと考えられる．そこで，主題語とより関係の強い兄弟語を抽出するため，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する．次に取得した兄弟語と主題語の共通する上位概念の重みをその兄弟語の重みとして全て加算することで兄弟語の重みを求める．本研究では兄弟語の重み上位 10 件を対立語候補として抽出する．

式 (3) に主題語 s のある上位概念 U_s の重み $Sta(U_s)$ を示す．ここでは，上位概念 U_s が持つ下位概念数 n と主題語 s のすべての兄弟概念数 N_s を用いて以下の式で求める．

$$Sta(U_s) = -\frac{\log n}{N_s} \tag{3}$$

式 (3) を用いて「又吉直樹」の上位概念の重みを計算した値を表 12 に示す．下位概念数が少ない上位概念ほど高い重みが与えられるのがわかる．

表 12: 主題語「又吉直樹」の上位概念とその重み

上位概念	下位概念数	上位概念の重み
お笑い芸人	1432	7.624185
日本のタレント	2317	7.1429843
ピース	17	12.057799
日本の小説家	2236	7.17856902
芥川賞受賞者	147	9.9005802
俳人	218	9.50651779
吉本興業	70	10.64251761
大阪府出身の人物	4626	6.45156501
1980 年生	3522	6.7242285
存命人物	72774	3.69589882

ここで得られた上位概念の重みを用いて，兄弟語の重みを求める．ある兄弟語 b について主題語と共通する上位概念に対して式 (3) を用いて重みを求め，以下の式 (4) を用いて兄弟語 b の重み $Rel(b)$ を求める．

$$Rel(b) = \sum_{i=0}^{n} Sta(Ub_i) \tag{4}$$

ここで，Ub は兄弟語 b と主題語の共通する上位概念を示し，n は兄弟語 b と主題語の共通する上位概念の総数を，$Sta(Ub_i)$ は兄弟語 b が持つある上位概念 Ub_i の重みを示す．

「又吉直樹」の兄弟語の重みについて式 (4) を用いて計算した結果の一部を表 13 に示す．これにより,「又吉直樹」の対立語は「綾部祐二」,「塚地武雅」,「岩尾望」,「大上邦博」,「前田登」,「安達健太郎」,「田中卓志」,「山根良顕」,「蛍原徹」,「松本康太」が対立語候補として得られる．

6.1.3 対立語の抽出

次に取得した対立語の候補から対立語を抽出するために語の認知度を求める．我々の提案する語の認知度は，検索結果数が近い語ほど認知度が近いと考え，ある語をクエリとした Google 検索結果数をその語の認知度とする．つまりは，主題語と対立語の候補それぞれの検索結果数を各々の語の認知度として取得する．認知度を取得後，式 (5) に示すように，主題語と対立語の候補の認知度の比率を求め，その比率が大きい対立語の候補上位 3 件を対立語とする．

$$Con(s, c) = 1 - \frac{\log\{|Cog(s) - Cog(c)|\}}{\max\{Cog(s), Cog(c)\}} \tag{5}$$

ここで，s は主題語を示し，c は対立語候補を示す．また $Cog(s)$ は主題語 s の検索結果数，$Cog(c)$ は対立語候補 c の検索結果数を示す．

表 13: 主題語「又吉直樹」の兄弟語とその重みの一部

兄弟語	重み
綾部祐二	39.105586
塚地武雅	30.937196
岩尾望	30.937196
大上邦博	30.937196
前田登	25.7160355
安達健太郎	25.7160355
田中卓志	25.605560
山根良顕	25.605560
蛍原徹	25.605560
松本康太	25.605560
徳井義実	25.605560
木村明浩	25.605560
桜田淳子	25.459220
月亭八光	25.459220
川島明	25.3270758

「又吉直樹」の対立語の候補となる語 10 件における認知度の対比率の結果を表 14 に示す．この結果より主題語「又吉直樹」に対して「綾部祐二」,「田中卓志」,「蛍原徹」が．対立語として得られる．

表 14: 主題語「又吉直樹」の対立語候補とその認知度

対立語候補	認知度
綾部祐二	0.5732394
田中卓志	0.3050409
蛍原徹	0.2374155
塚地武雅	0.2201494
岩尾望	0.0964093
山根良顕	0.0820209
松本康太	0.0238199
木村明浩	0.0148991
安達健太郎	0.0118776
前田登	0.0101510

6.2 暗黙的観点に基づく対立記事抽出のための対立語抽出手法

本研究では，観点毎に対立記事は異なる．また，暗黙的観点は複数存在するため，暗黙的観点毎に対立記事も異なると考えられる．そこで，それぞれの暗黙的観点から対立語の抽出を行う．暗黙的観点に基づく対立語抽出手法は，トピック抽出に基づく暗黙的観点の抽出手法と概念構造に基づく暗黙的観点の抽出手法でそれぞれ異なる．

6.2.1 トピック抽出に基づく暗黙的観点の対立語候補の抽出

例えば5.2.1 節で得られた暗黙的観点は語の上位概念であるとは限らないため，暗黙的観点と概念辞書を用いて対立語を抽出することは困難である．この時，本研究の暗黙的観点の定義より主題語と暗黙的観点は関係が強い語であるため，対立語と暗黙的観点も関係の強い語であると考えられる．そのため，暗黙的観点となる語を含むニュース記事の主題となる語は暗黙的観点との関係が強いと考えられる．そこで，このような語を対立語候補として抽出し，これら対立語候補から対立語の抽出を行う．

以下と図 4 にトピック抽出に基づく暗黙的観点の対立語抽出手法の概要を示す．

(1) 暗黙的観点を用いてニュース記事群を取得する．

(2) それぞれのニュース記事の主題語を抽出する．

(3) 概念辞書を用いてそれぞれの主題語の上位概念を取得する．

(4) 閲覧記事の主題語の上位概念を取得する．

(5) (3) と (4) で得られた上位概念の類似度計算を行う．

(6) (5) の結果より類似度の高い主題語を対立語候補として抽出する．

(7) 対立語候補と主題語との認知度計算を行い対立語を決定する．

具体的には，まずある暗黙的観点を含むニュース記事群 100 件を取得する．この時，ニュース記事の Web ページのみを取得するため，ニュースサイトのドメインを用いてサイト指定検索を行う．次に得られたニュース記事群それぞれから式 (1) を用いて主題語を抽出する．この時，暗黙的観点を含むニュース記事は閲覧記事と内容が大きく異なる場合がある．このような場合，得られる主題語は閲覧記事の主題語との関係が弱いと考えられ，対立語として抽出することは不適切である．そこでニュース群から得られた主題語の中で閲覧記事の主題語と関係の強い語を抽出するため，ニュース群から得られた主題語の上位概念と閲覧記事の主題語の上位概念の類似度計算を行い，類似度の高い主題語を対立語候補として抽出する．この時，類似度計算にはダイス係数を用いる．ニュース群から得られたある主題語 r と閲覧記事の主題語 s の共通する上位概念の数を $|Ur, s|$，ニュース群から得られたある主題語 r の上位概念の数を $|U_r|$，閲覧記事の主題語 s の上位概念の数を $|U_s|$ とし，ニュース群から得られたある主題語 r と閲覧記事の主題語 S の類似度 $Cos(S, r)$ を式 (6) に示す．この類似度 $Cos(S, r)$ が高い上位 10 件の語を対立語候補として抽出する．このようにして，すべての暗黙的観点毎に対立語候補を抽出する．

$$Cos(s, r) = \frac{|U_{r,s}|}{|U_s||U_r|} \tag{6}$$

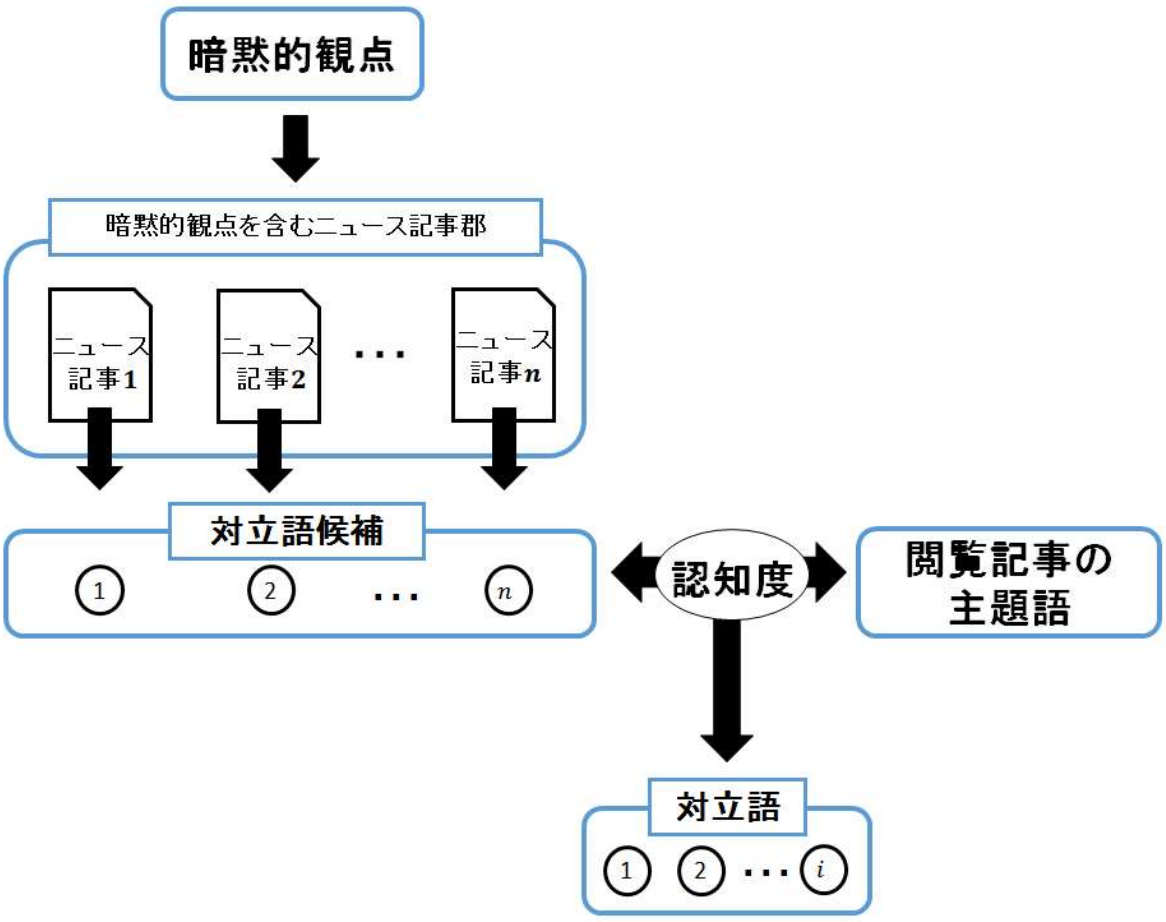

図 4: トピック抽出に基づく暗黙的観点の対立語抽出の概要

例えば，表 6 より得られたトピック抽出に基づく暗黙的観点「小説」から抽出されるニュース群から得られた主題語の一部と主題語との類似度を表 15 に表す．この結果より,「山下澄人」,「吉田修一」,「ビートたけし」,「高井有一」,「つぶやきシロー」,「村上春樹」,「塩田武士」,「山本一力」,「恩田陸」,「藤沢周平」が対立語候補として抽出される．

表 15: 暗黙的観点「小説」から抽出された主題語の一部とその重み

主題語	閲覧記事の主題語との類似度
山下澄人	20.77504812
吉田修一	20.77504812
ビートたけし	18.46306867
高井有一	17.07914929
つぶやきシロー	11.32008434
村上春樹	10.87446785
塩田武士	10.87446785
山本一力	10.87446785
恩田陸	10.87446785
藤沢周平	7.17856902
和田裕美	3.69589883
ピクシブ	0
サッカー	0
幻冬舎	0

6.2.2 トピック抽出に基づく暗黙的観点の対立語の抽出

次に，ここで得られた対立語候補から対立語の抽出を行う．対立語の定義より，対立語と閲覧記事の主題語は同程度の認知度を持つと考えられる．そこで式 (4) を用いて対立語候補と閲覧記事の主題語との認知度計算を行う．この認知度比が高い上位 3 件の語を対立語として決定する．

表 15 より得られた対立語候補と閲覧記事の主題語「又吉直樹」との認知度比を表 16 に示す．この結果より，暗黙的観点「小説」における対立語は「恩田陸」,「山下澄人」,「吉田修一」となる．

6.2.3 概念構造に基づく暗黙的観点の対立語の抽出

例えば 5.2.2 節で得られた暗黙的観点「お笑い芸人」は「又吉直樹」の上位概念である．そこで対立語を抽出するために「又吉直樹」の上位概念が持つ全ての下位概念語を兄弟語とするのではなく，暗黙的観点が持つ下位概念語を兄弟語として抽出を行い，対立語の抽出を行う．同様にして，5.2.2 節で得られた全ての暗黙的観点から観点毎に対立語を抽出する．

以下に概念構造に基づく暗黙的観点の対立語抽出手法の概要を示す．

(1) 暗黙的観点が持つ下位概念語を兄弟語として取得する．

表 16: 暗黙的観点「小説」から抽出された対立語候補とその認知度

対立語候補	認知度比
恩田陸	0.739396
山下澄人	0.494967
吉田修一	0.278422
ビートたけし	0.23089
村上春樹	0.1349522
つぶやきシロー	0.130941
和田裕美	0.067920
山本一力	0.057201
塩田武士	0.031877
高井有一	0.016409

(2) 兄弟語が持つ上位概念の重み計算を行う．

(3) 兄弟語から対立語候補を抽出する．

(4) 対立語候補から認知度を用いて対立語を抽出する．

具体的には，まず暗黙的観点が持つ下位概念語をすべて取得し，兄弟語とする．次に 6.1.2 節と同様にしてそれぞれの兄弟語が持つ上位概念の重みの計算を行い，兄弟語の重みを決定する．この兄弟語の重みが高い上位 10 件の語を対立語候補として抽出する．さらに 6.2.2 と同様にして対立語候補と閲覧記事の主題語との認知度比を求め，認知度比が上位 3 件の語を対立語として抽出する．表 6 より得られた概念構造に基づく暗黙的観点「お笑い芸人」から抽出された対立語候補と「又吉直樹」との認知度比を表 17 に示す．この結果より，暗黙的観点「お笑い芸人」における対立語は「星野卓也」,「星ルイス」,「西野亮廣」となる．

表 17: 暗黙的観点「お笑い芸人」から抽出された対立語候補とその認知度

対立語候補	認知度比
星野卓也	0.945026
星ルイス	0.943209
西野亮廣	0.798429
西國坊明學	0.712041
末吉くん	0.712041
ますだおかだ	0.678010
清水アキラ	0.639865
西本はるか	0.639866
綾部祐二	0.573944
村田渚	0.518324607
西川きよし	0.492146597

7　対立記事の抽出手法

本研究における対立記事の定義は以下のとおりである．

1. 閲覧記事の主題語と対立する語に関する記事
2. 閲覧記事と類似する内容の記事

ここで，閲覧記事の主題語と対立する語は本研究での対立語を指す．また，対立記事は観点毎に異なる場合がある．そこで，本研究では観点毎に対立記事の抽出を行う．

7.1　対立記事候補の抽出

対立記事は対立語を含む記事である．そこで対立語をクエリとして得られるニュース記事群を対立記事候補として抽出を行う．しかしながら，対立記事の存在する対立語に対して，その対立語のみをクエリとしてニュース記事を取得する場合，対立語を含む対立記事よりも新しい記事が大量に存在すると対立記事が抽出することが困難であると考えられる．この時，対立記事の定義より，対立記事は閲覧記事と類似しているため，対立記事閲覧記事に記載されている重要な単語を含む場合が多いと考えられる．そこで対立記事候補を取得するため対立語と閲覧記事の明示的観点を用いて対立記事候補の抽出を行う．この時，対立語や閲覧記事の明示的観点は複数抽出されるため，クエリとしてある明示的観点に対して 1 つの対立語の組み合わせを決定し，対立記事候補として抽出する．この時，明示的観点と対立語との組み合わせを決めるために，明示的観点毎に式 (2) を用いて共起度計算を行い，最も値の高い組み合わせの対立語を決定する．5.1 節で得られた明示的観点「火花」,「コメント」,「たくさん」,「本」,「夏」,「突破」,「芥川賞」と 6.2.2 節で得られた明示的観点の対立語「田中卓志」,「蛍原徹」,「綾部祐二」の共起度計算の結果を表 18 に示す．この結果より，明示的観点「火花」には対立語「田中卓志」,「コメント」には「田中卓志」,「たくさん」には「蛍原徹」,「本」には「田中卓志」,「夏」には「田中卓志」,「突破」には「田中卓志」,「芥川賞」には「綾部祐二」の組み合わせとなる．

表 18: 明示的観点と対立語との共起度

明示的観点	対立語	共起度	明示的観点	対立語	共起度
火花	田中卓志	0.0936	コメント	田中卓志	0.0114
	蛍原徹	0.0808		蛍原徹	0.0027
	綾部祐二	0.0063		綾部祐二	0.0018
たくさん	蛍原徹	0.0336	本	田中卓志	0.0005
	田中卓志	0.0193		綾部祐二	0.0003
	綾部祐二	0.0014		蛍原徹	0.0003
夏	田中卓志	0.0141	突破	田中卓志	0.0229
	綾部祐二	0.0084		綾部祐二	0.0013
	蛍原徹	0.0052		蛍原徹	0.0012
芥川賞	綾部祐二	0.0237			
	蛍原徹	0.0141			
	田中卓志	0.0119			

ここで，上記のように明示的観点に基づく対立記事を抽出するために明示的観点毎に対立語の組み合わせを決定したが，暗黙的観点に基づく対立記事を抽出する際は暗黙的観点毎に対立記事を抽出する．この時，暗黙的観点から抽出した対立語の方が明示的観点より暗黙的観点との関係が強いため，暗黙的観点に基づく対立記事の候補を抽出するために，対立語毎に明示的観点との組み合わせを決定し，この組み合わせによって取得するニュース群をその暗黙的観点の対立記事候補とする．6.2.1 節より得られたトピック抽出に基づく暗黙的観点のうち「小説」における対立語「恩田陸」,「山下澄人」,「吉田修一」と明示的観点との共起度と，6.2.3 節より得られた概念構造に基づく暗黙的観点のうち「お笑い芸人」における対立語「星野卓也」,「星ルイス」,「西野亮廣」と明示的観点との共起度を表 19 に示す．この結果より，トピック抽出に基づく暗黙的観点の対立語「恩田陸」には明示的観点「芥川賞」,「山下澄人」には「芥川賞」,「吉田修一」には「火花」という組み合わせとなり，概念構造に基づく暗黙的観点の対立語「星野卓也」には明示的観点「火花」,「星ルイス」には「夏」,「西野亮廣」には「突破」という組み合わせとなる．

表 19: 暗黙的観点の対立語と明示的観点との共起度

トピック抽出に基づく暗黙的観点「小説」の対立語	明示的観点	共起度	概念構造に基づく暗黙的観点「お笑い芸人」の対立語	明示的観点	共起度
恩田陸	芥川賞	0.3279	星野卓也	火花	0.1221
	たくさん	0.0729		突破	0.0233
	夏	0.0339		たくさん	0.0197
	コメント	0.0116		夏	0.0142
	火花	0.0104		芥川賞	0.0138
	突破	0.0043		コメント	0.0069
	本	0.0017		本	0.0008
山下澄人	芥川賞	0.5036	星ルイス	夏	0.0216
	突破	0.0374		芥川賞	0.0167
	夏	0.0229		コメント	0.0129
	たくさん	0.0188		火花	0.0057
	火花	0.0082		たくさん	0.0029
	コメント	0.0023		突破	0.0020
	本	0.0012		本	0.0008
吉田修一	火花	0.2089	西野亮廣	突破	0.1780
	突破	0.0415		火花	0.0342
	たくさん	0.0353		たくさん	0.0288
	芥川賞	0.0353		芥川賞	0.0091
	夏	0.0255		夏	0.0066
	コメント	0.0131		コメント	0.0046
	本	0.0013		本	0.0006

7.2 対立記事の抽出

観点毎に得られた対立記事候補から対立記事の抽出を行う．この時，対立記事の定義より，閲覧記事と対立記事は類似している．そこで閲覧記事に出現する名詞とある対立記事候補に出現する名詞を用いて類似度計算を行い，類似度の最も高い記事を対立記事に決定する．類似度計算にはコサイン類似度を用いる．このようにして，すべての観点に対して対立記事を抽出し，提示する．

ここで，表 6 より得られた明示的観点のうち「芥川賞」から抽出される対立記事，トピック抽出に基づく暗黙的観点のうち「小説」から抽出される対立記事，概念構造に基づく暗黙的観点のうち「お笑い芸人」から抽出される対立記事のタイトルを表 20 に示す．

表 20: 「又吉直樹」の記事から抽出された対立記事のタイトル

観点	対立記事のタイトル
明示的観点 「芥川賞」	相方の綾部さん「アシスタントになる覚悟できた」 同居のパンサー向井さん「お母さんのよう」
トピック抽出による 暗黙的観点「小説」	芥川賞・山下澄人さん「しんせかい」／直木賞・ 恩田陸さん「蜜蜂と遠雷」
概念構造による 暗黙的観点 「お笑い芸人」	キングコング・西野亮廣の作品が１０００万円で 売れた！　高野山三宝院に奉納へ

8 評価実験

本研究では対立記事抽出の有用性を計るため，主題語の抽出，観点の抽出，対立語の抽出，そして対立記事の抽出について評価実験を行った．

8.1 主題語抽出の評価実験

本研究では，まずニュース記事から主題語が正しく抽出できているかを計るため，産経ニュースからランダムに選んだニュース記事 20 件を評価実験を行った．20 件のニュース記事のタイトルを表 21 に示す．ニュース記事 20 件を用いて評価方法は，8 名の被験者にニュース記事から抽出された主題語がその記事の主題語と言えない場合は 1，適切である場合は 5 として 5 段階評価で判断を行う．抽出された主題語に対して 8 名の被験者による評価の平均値をその主題語の評価値とする．

評価実験により得られた 20 件の主題語の評価値の割合を図 5 に示す．この結果より，評価値が 3.0 以上となった割合が 95 %であり，主題語の抽出はほぼ正確であることがわかる．

この時,「鹿児島県知事、川内原発再稼働に同意を表明」という記事で抽出された主題語は「鹿児島県」となり，評価値の平均は 3.57 となった．評価が低かった原因として，このニュース記事の主題語は「鹿児島県」,「鹿児島県知事」,「川内原発」の 3 語考えられ，被験者によって判断が分かれたためであると考えられる．また，このニュース記事はタイトルから「鹿児島県知

表 21: 評価実験で用いた記事 20 件のタイトル

ニュース記事のタイトル
関電、中間黒字と 26 億円を確保　通期は「原発ゼロなら赤字転落」と八木社長
ダイキン、省エネエアコンの研究開発へ　小惑星探査機「はやぶさ」の技術を応用
宮崎監督にアカデミー名誉賞授与　黒澤明監督依頼の快挙...「死ぬまでアニメ作る」
イズミヤ統合で過去最高益　H2O の中間決算 増税影響も想定より少なく
鹿児島県知事、川内原発再稼働に同意を表明
クボタが決算期を変更　海外事業の拡大に対応
ＡＣミランが２３歳ＦＷ獲得へ
ＢＳＥ発生、ノルウェー産牛肉輸入停止　厚労省発表
ロナウド氏、現役復帰か　共同オーナー就任の米ＭＬＳ下部リーグで
東ガスが再生エネ電力を購入　千葉県の風力１万２千キロワット分
アサヒビールが糖質ゼロの第３のビールを発売　度数６％と両立
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
オバマ米大統領が印パレードに出席、米大統領で初主賓
オバマ氏、会談拒否か　ネタニヤフ首相の訪米時
高倉健さんしのぶ展示会が人気　１３０点以上の資料　北九州市
ドイツ、１月の物価０・３％下落　５年４カ月ぶりにマイナス
大ガス決算、純利益３・３倍　株売却、前年損失の反動も
パナ、自販機生産態勢を見直し　業務用強化で群馬に
京阪電鉄が持ち株会社に移行　来年４月、グループ強化へ
ＶＩＰ会見にイチロー「ただただ恐縮している」

事が同意を表明した」という出来事と「川内原発が再稼働する」という出来事という 2 つの出来事を表している．そのため，評価値が低くなったと考えられる．これに対して，抽出された主題語「鹿児島県」に対して高い評価を行った被験者も存在する．この理由として「鹿児島県知事」は「鹿児島県」の代表であることから「鹿児島県知事」と「鹿児島県」は同じと判断し，高い評価を行ったのではないかと考えられる．

8.2　観点抽出の評価実験

本研究では，ニュース記事から観点が正しく抽出できているかを計るために評価実験を行った．

この時，観点は明示的観点，トピック抽出に基づく暗黙的観点，概念構造に基づく暗黙的観点でそれぞれ異なるため，これらについてそれぞれの評価を行う．

8.2.1　明示的観点の評価実験

明示的観点が正しく抽出されているかを評価するため，8.1 節と同様に表 21 に示すニュース記事 20 件に対してそれぞれ抽出された明示的観点を 5 段階評価で実験を行った．この時，抽出さ

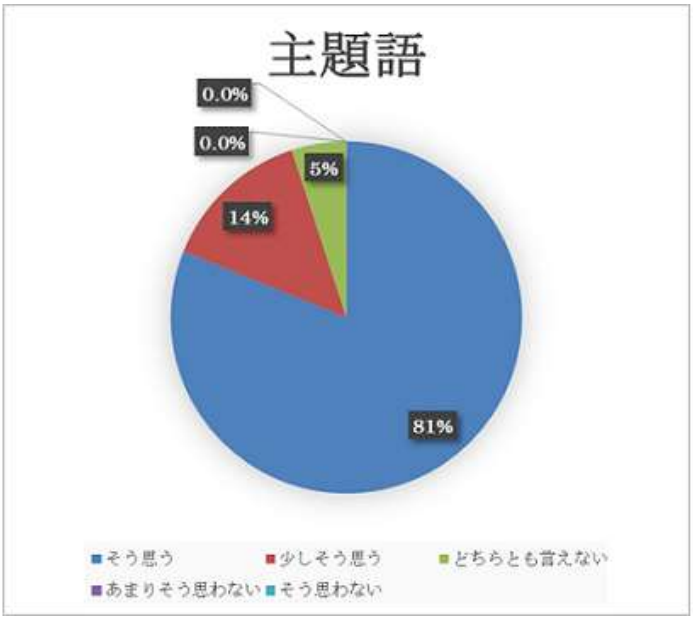

図 5: 主題語抽出の評価の割合

れた明示的観点それぞれに対し 8 名の被験者の評価の平均値をその明示的観点の評価値とする．

抽出された明示的観点すべての評価値の割合を図 6 に示す．この結果より，評価値が 3.0 以上となった割合が 56 %となった．

この時，抽出された明示的観点の数は 20 件のニュース記事すべて合わせて 119 語であった．このうち,「確保」や「発表」などサ変活用の名詞が 44 語存在し，このうち 33 語が評価値 3.0 以下であることが分かった．そのため，明示的観点の抽出においてサ変活用の名詞を除去することでより正確に明示的観点の抽出ができると考えられる．

8.2.2 トピック抽出に基づく暗黙的観点の評価実験

本研究では，トピック抽出に基づく暗黙的観点が正しく抽出できているかを計るため，暗黙的観点の抽出とその評価を行った．実験に用いるニュース記事のタイトルを表 22 に示す．これらの記事を用いてトピック抽出を行い暗黙的観点を抽出し，暗黙的観点が正しく抽出出来ているかの評価を行った．

また表 22 のニュース記事それぞれから抽出される主題語を表 23，明示的観点を表 24 に示す．

ここで，記事 1 からは主題語「阿蘇山」，記事 2 からは主題語「任天堂」，記事 3 からは主題語「日立製作所」，記事 4 からは主題語「イチロー」を用いてそれぞれのトピック抽出を行った．抽出されたトピックとそのトピックを持つクラスタに含まれているニュース記事数の一部を表 25 に示す．

表 25 よりトピックに基づく暗黙的観点は，記事 1 の主題語「阿蘇山」に対して「地震」,「火山」が抽出される．この時，阿蘇山は活火山であるため暗黙的観点として「火山」は適切であると考えられる．しかしながら「地震」については熊本地震による影響でトピックとして抽出

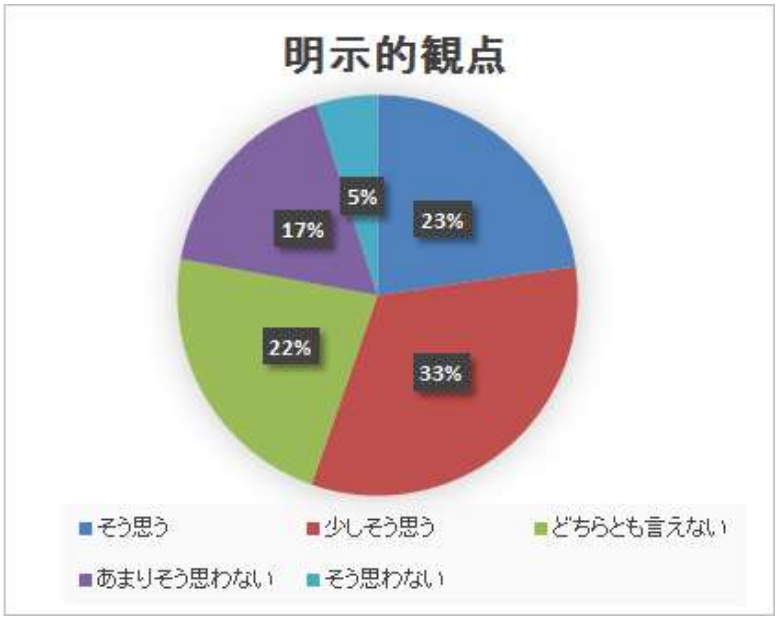

図 6: 明示的観点の評価の割合

表 22: 評価実験で用いた記事 4 件のタイトル

記事番号	ニュース記事のタイトル
記事 1	阿蘇山で爆発的噴火　警戒レベルを 2 から 3 に引き上げ
記事 2	任天堂、「Ｗｉｉ U」の生産を年内にも終了へ　次世代機に注力
記事 3	日立が「レンズなし」のカメラ開発　フイルムでレンズ代用、薄く軽く低価格に
記事 4	【イチロー 3 千安打】三塁打で達成は 2 人目　９６年のモリター以来

表 23: 評価実験で用いた記事 4 件から抽出される主題語

記事番号	主題語	記事番号	主題語
記事 1	阿蘇山	記事 2	任天堂
記事 3	日立製作所	記事 4	イチロー

されたと考えられるが，熊本地震が発生する以前は九州地方において地震の発生がほとんどないため阿蘇山の暗黙的観点としては不適切であると考えられる．記事 2 の主題語「任天堂」において抽出される暗黙的観点は「人気」,「ゲーム」,「終値」,「赤字」となる．任天堂は玩具もしくはゲーム会社という側面が強い．そのため「ゲーム」は暗黙的観点として適切であると考えられる．また，任天堂は 2013 年から 2015 年まで売上高が低下し続けていることから赤字というトピックが抽出されたと考えられる．しかしながら赤字や黒字といった状態は変動しやすく，

表 24: 評価実験で用いた記事 4 件から抽出される明示的観点

記事番号	明示的観点	記事番号	明示的観点
記事 1	噴火	記事 2	生産
	爆発		終了
	警戒		年内
	火口		
	引き上げ		
	レベル		
記事 3	カメラ	記事 4	達成
	技術		安打
	開発		コーチ
			打撃
			メジャー
			通算

表 25: 抽出されたトピックとそのトピックを持つクラスタに含まれる記事数

記事 1		記事 2		記事 3		記事 4	
トピック	記事数	トピック	記事数	トピック	記事数	トピック	記事数
地震	37	人気	22	社長	14	野球	26
火山	32	ゲーム	18	効率	13	打率	15
降灰	9	終値	18	社員	11	安打	12
マグマ	6	赤字	12	業績	11	選手	10
断層	6	次世代	9	技術	9	メジャー	8
火口	5	ソフト	6	レンズ	8	通算	7
不安	4	Wii	6	ガラス	6	史上	7
状況	4	ゲーム機	5	大手	6	大台	3

任天堂の持つ側面として赤字企業であるとは言えない．そのため「赤字」は暗黙的観点として不適切である.「人気」は任天堂ではなく任天堂が生産している商品に関する観点であると考えられ，任天堂の暗黙的観点としては不適切である.「終値」については任天堂が株式会社であるため株価変動に関するニュース記事が多く，トピックとして抽出されたと考えられる．しかしながら任天堂の側面として「終値」を持つとは言えない．そのため暗黙的観点として不適切であると考えられる．記事 3 の主題語「日立製作所」に対して抽出される暗黙的観点は「社長」,「効率」,「社員」,「業績」となる．日立製作所に関するニュースにおいて社長の発言が取り上げられる場合が多い．そのため「社長」という語がトピックとして抽出されたと考えられる．しかしながら日立製作所の持つ側面として「社長」という語が適切であるとは言えないため，暗黙的観点として不適切であると考えられる．同様にして「効率」,「社員」,「業績」も暗黙的観点として不適切であると考えられる．記事 4 は主題語「イチロー」に対して「野球」,「打率」,「安打」,「選手」となる．イチローは野球選手であるため,「野球」,「選手」は暗黙的観点として

適切であると考えられる．また，記事 4 に記載されているようにイチローの安打数は野球界においてトップクラスであり,「安打」や安打に関連する「打率」も暗黙的観点として適切であると考えられる．

この結果より，抽出された暗黙的観点 14 語のうち暗黙的観点として適切な語は 6 語となり，適合率は約 43 %であった．また，主題語によっては適切な暗黙的観点が 1 語も抽出できない場合があることが分かった．

8.2.3 概念構造に基づく暗黙的観点の評価実験

本研究で提案する概念構造に基づく暗黙的観点が正しく抽出できているかを計るため，暗黙的観点の抽出とその評価を行った．暗黙的観点の抽出には，トピック抽出に基づく暗黙的観点の抽出と同様に表 22 のニュース記事 4 件を用いる．

記事 1 からは主題語「阿蘇山」，記事 2 からは主題語「任天堂」，記事 3 からは主題語「日立製作所」，記事 4 からは主題語「イチロー」を用いてそれぞれ概念構造に基づく暗黙的観点の抽出を行った．抽出された主題語の上位概念と主題語との共起度を表 26〜28 に示す．

表 26: 主題語から抽出した上位概念と共起度 1

「阿蘇山」の上位概念	共起度	「任天堂」の上位概念	共起度
熊本県の自然景勝地	0.08812	日本の玩具メーカー	0.07100
阿蘇市	0.04647	京都市南区の企業	0.06573
九州地方の火山	0.02778	1947 年設立の企業	0.05638
日本の火山災害	0.02684	日本のコンピュータゲームメーカー・ブランド	0.05541
日本百名山	0.01371	東証一部上場企業	0.04774
熊本県の山	0.01356	多国籍企業	0.03144
地質・鉱物天然記念物	0.0823	東山区の歴史	0.01521
熊本県にある国指定の名勝	0.0793	老舗企業 (明治創業)	0.01484

表 27: 主題語から抽出した上位概念と共起度 2

「日立製作所」の上位概念	共起度
日本の情報・通信業	0.07393
日本の鉄道車両メーカー	0.04726
日本の輸送機器メーカー	0.04224
日立グループ	0.03383
日本の電気機器メーカー	0.02807
1920 年設立の企業	0.01626
名証一部上場企業	0.01227
自動車燃料供給装置メーカー	0.01136
老舗企業 (明治創業)	0.01123

表 28: 主題語から抽出した上位概念と共起度 3

「イチロー」の上位概念	共起度
MLB の日本人選手	0.04496
愛知県出身の人物	0.03706
日本の野球選手	0.03450
首位打者 (NPB)	0.03158
盗塁王 (MLB)	0.02675
ワールド・ベースボール・クラシック日本代表選手	0.02457
オリックス・バファローズ及びその前身球団の選手	00.2117
打点王 (NPB)	0.01920

表 26〜28 より，記事 1 の主題語「阿蘇山」に対する暗黙的観点は「熊本県の自然景勝地」,「阿蘇市」,「九州地方の火山」,「日本の火山災害」となる．ここで，主題語である「阿蘇山」は地名を表している．そのため地名情報である「熊本県の自然景勝地」,「阿蘇市」は暗黙的観点として適切であると考えられる．また,「九州地方の火山」は阿蘇山が活火山であることから暗黙的観点として適切である．さらに「日本の火山災害」は阿蘇山が 2015 年にも噴火し大きな災害となったことから，暗黙的観点として適切であると考えられる．記事 2 の主題語「任天堂」に対する暗黙的観点は「日本の玩具メーカー」,「京都市南区の企業」,「1947 年設立の企業」,「日本のコンピュータゲームメーカー・ブランド」,「東証一部上場企業」,「多国籍企業」となる.「日本の玩具メーカー」や「日本のコンピュータゲームメーカー・ブランド」,「東証一部上場企業」は任天堂が持つ側面と考えられる．そのため暗黙的観点として適切である．また，任天堂は世界的に有名であり，海外からの労働者が存在すると容易に考えられる．そのため「多国籍企業」は暗黙的観点として適切であると考えられる．一方,「京都市南区の企業」は京都市南区という非常に狭い地域に限定されており，多くの人々が知っている情報でないと考えられ，暗黙的観点として適切でない．また，同様にして「1947 年設立の企業も暗黙的観点として適切でないと考えられる．記事 3 の主題語「日立製作所」に対する暗黙的観点は「日本の情報・通信業」,「日本の鉄道車両メーカー」,「日本の輸送機器メーカー」,「日立グループ」,「日本の電気機器メーカー」となる．ここで抽出された語のうち「日本の情報・通信業」,「日本の鉄道車両メーカー」,「日本の輸送機器メーカー」,「日立グループ」,「日本の電気機器メーカー」は日立製作所の側面を表す語であると考えられる．そのため，暗黙的観点として適切である．記事 4 の主題語「イチロー」に対する暗黙的観点は「MLB の日本人選手」,「愛知県出身の人物」,「日本の野球選手」,「首位打者 (NPB)」,「盗塁王 (MLB)」,「ワールド・ベースボール・クラシック日本代表選手」,「オリックス・バファローズ及びその前身球団の選手」となる．抽出されたこれらの語のうち「MLB の日本人選手」,「愛知県出身の人物」,「日本の野球選手」,「ワールド・ベースボール・クラシック日本代表選手」,「オリックス・バファローズ及びその前身球団の選手」は多くの人々が知っている情報であり，暗黙的観点として適切であると考えられる．しかしながら「首位打者 (NPB)」,「盗塁王 (MLB)」は野球のシーズン毎に変化するため，常にイチローが持つ側面であるとは言えない．そのため暗黙的観点として適切でないと考えられる．

この結果より，抽出された暗黙的観点 22 語のうち暗黙的観点として適切な語は 18 語となり，適合率は約 81 %となる．そのためトピック抽出に基づく暗黙的観点より概念構造に基づく暗黙的観点の方が適切な暗黙的観点を抽出できると考えられる．しかしながら，トピック抽出に基

図 7: 明示的観点に対する対立語の評価の割合

づく暗黙的観点で抽出した語のうち暗黙的観点に適切であると考えられる語のうち概念情報でない語も存在する．そのため，概念構造に基づく暗黙的観点では抽出できない暗黙的観点が存在することが分かった．

8.3 対立語抽出の評価実験

本研究で提案する対立語の抽出手法によって正しく抽出できているかを計るために評価実験を行った．

この時，明示的観点に対する対立語，トピック抽出に基づく暗黙的観点に対する対立語，概念構造に基づく暗黙的観点に対する対立語で抽出手法がそれぞれ異なるため，これらについてそれぞれの評価を行う．

8.3.1 明示的観点に対する対立語の評価実験

明示的観点に対する対立語が正しく抽出されているかを評価するため，8.2.1 節と同様に表 21 に示すニュース記事 20 件に対してそれぞれ抽出された対立語を 5 段階評価で実験を行った．この時，抽出された対立語それぞれに対し 8 名の被験者の評価の平均値をその対立語の評価値とする．

抽出された対立語の評価値の割合を図 7 に示す．この結果より，評価値が 3.0 以上となった割合が 54 %となった．

抽出結果のうち，評価の低かった例としては「関西電力」や「東京ガス」の対立語としてどちらも「東芝」という語が抽出された．原因として「東芝」が持つ上位概念を 24 語持ち,「関西

電力」や「東京ガス」と共通する上位概念が多く存在したためであると考えられる．このように共通する上位概念数が多いだけで対立語として抽出されてしまう語が多くあったため，評価値が低くなってしまったと考えられる．

8.3.2 トピック抽出に基づく暗黙的観点に対する対立語の評価実験

トピック抽出に基づく暗黙的観点に対する対立語が正しく抽出できているかを計るため，対立語の抽出とその評価を行った．対立語の抽出には暗黙的観点の抽出と同様に表 22 のニュース記事 4 件を用いる．

抽出された対立語のうち，8.2.1 節において暗黙的観点が正しく抽出されたと判断した記事 1 の「火山」と記事 4 の「打率」におけるそれぞれの対立語候補と主題語との認知度を表 29 に示す．

表 29: 暗黙的観点「火山」,「打率」の対立語候補と主題語との認知度

暗黙的観点	対立語候補	認知度	暗黙的観点	対立語候補	認知度
火山	伊豆大島	0.1572	打率	T-岡田	0.4944
	蔵王山	0.1427		山田哲人	0.3253
	桜島	0.1237		田村龍弘	0.1138
	富山地方気象台	0.0887		中井康之	0.0076
	えびの高原	0.0821		大谷翔平	0.0011
	三瓶山	0.0722		鈴木誠也	0.0010
	鹿児島県	0.0477		三浦大輔	0.0008
	薩摩硫黄島	0.0405		新井貴浩	0.0004
	西之島	0.0394		マット・マートン	0.0004

表 29 より，記事 1 の「火山」に対する対立語は「伊豆大島」,「蔵王山」,「桜島」となる.「伊豆大島」は東京都に属する島の名称である．しかしながら伊豆大島は水深 300〜400 メートルほどの海底からそびえる活火山の陸上部分である．しかしながら，伊豆大島は過去 30 年間噴火をしておらず,「阿蘇山」の対立語として適切でないと考えられる.「蔵王山」は宮城県と山形県にまたがる山を指す.「蔵王山」も活火山であり，2015 年に噴火の可能性があるとして火口周辺警報が発令されたことから「阿蘇山」の対立語として適切であると考えられる.「桜島」は鹿児島県に属する島の名称である.「桜島」も活火山であり 2016 年に噴火している．そのため,「阿蘇山」の対立語として適切であると考えられる．記事 4 の「打率」に対する対立語は「T-岡田」,「山田哲人」,「田村龍弘」となる.「T-岡田」,「山田哲人」,「田村龍弘」は全てプロ野球選手という点でイチローと一致する．また,「T-岡田」はイチローと同様に外野手であることから対立語として適切であると考えられる．また「山田哲人」は内野手であるが年棒が 2.2 億円でありイチローの年棒 200 万ドルと近いことから対立語として適切であると考えられる．

ここで，記事 2 の主題語「任天堂」において暗黙的観点として適切であると判断した「ゲーム」から抽出された対立語は存在しなかった．原因として「ゲーム」をクエリとして得られるニュース記事の主題語のうち「任天堂」と一致する上位概念が存在しない語であったり，概念辞書に存在しない企業名であったためである．そのため，適切な暗黙的観点でも対立語が抽出できない場合があることが分かった．

8.3.3　概念構造に基づく暗黙的観点に対する対立語の評価実験

概念構造に基づく暗黙的観点に対する対立語が正しく抽出できているかを計るため，対立語の抽出とその評価を行った．対立語の抽出には 8.3.2 節と同様に表 22 のニュース記事 4 件を用いる．

抽出された対立語のうち，8.2.2 節において暗黙的観点が正しく抽出されたと判断した記事 1 の「九州地方の火山」，記事 2 の「日本の玩具メーカー」，記事 3 の「日本の電気機器メーカー」，記事 4 の「MLB の日本人選手」におけるそれぞれの対立語候補と主題語との認知度を表 30〜33 に示す．

表 30: 暗黙的観点「九州地方の火山」の対立語候補と主題語との認知度

暗黙的観点	対立語候補	認知度
九州地方の火山	黒島	0.652733
	金峰山	0.591133
	鶴見岳	0.591133
	小臥蛇島	0.585014
	硫黄島	0.564673
	由布岳	0.561576
	若尊	0.431915
	経ヶ岳	0.381773

表 31: 暗黙的観点「日本の玩具メーカー」の対立語候補と主題語との認知度

暗黙的観点	対立語候補	認知度
日本の玩具メーカー	バンダイ	0.973064
	クローバー	0.937297
	やまと	0.881199
	レゴ	0.476355
	ピープル	0.222607
	河田	0.192618
	タカラトミー	0.155709

表 30 より，暗黙的観点「九州地方の火山」から抽出される対立語は「黒島」,「金峰山」,「鶴見岳」となった．しかしながら「黒島」,「金峰山」,「鶴見岳」のうちどの対立語も過去に噴火や火山が活性化した記録が存在せず,「阿蘇山」の対立語として不適切であると考えられる．

次に，表 31 より，暗黙的観点「日本の玩具メーカー」から抽出される対立語は「バンダイ」,「クローバー」,「やまと」となった.「バンダイ」は玩具やゲームメーカーとして有名であり,「任天堂」の対立語として適切であると考えられる．しかしながら,「クローバー」や「やまと」は玩具メーカーであるが,「任天堂」との知名度に大きな差があると考えられる．また,「クローバー」や「やまと」は様々な語に用いられるため，認知度が高くなったと考えられる．

次に，表 32 より，暗黙的観点「日本の電気機器メーカー」から抽出される対立語は「新川」,「村田製作所」,「太陽誘電」となった.「新川」は半導体の製造装置のメーカーであり世界シェア

表 32: 暗黙的観点「日本の電気機器メーカー」の対立語候補と主題語との認知度

暗黙的観点	対立語候補	認知度
日本の電気機器メーカー	新川	0.741679
	村田製作所	0.521367
	太陽誘電	0.082264
	大東電機工業	0.081837
	小糸製作所	0.078418
	太洋工業	0.068162
	図研	0.051282

表 33: 暗黙的観点「MLB の日本人選手」の対立語候補と主題語との認知度

暗黙的観点	対立語候補	認知度
MLB の日本人選手	マック鈴木	0.335329
	高橋建	0.305389
	ダルビッシュ有	0.026570
	前田健太	0.02604
	松坂大輔	0.224850
	斎藤隆	0.222754
	松井秀喜	0.161077
	マイケル中村	0.16077

第 3 位の業績を持つ．しかしながら,「日立製作所」と同程度の知名度を持つとは言えない．そのため「日立製作所」の対立語として適切でないと考えられる.「村田製作所」は電子部品のメーカーとして世界トップクラスの実績を持つ企業である．そのため「日立製作所」の対立語として適切であると考えられる.「太陽誘電」は CD-R などの記録メディアのメーカーである．しかしながらこの語も「日立製作所」と同程度の知名度を持つとは言えない．そのため「日立製作所」の対立語として適切でないと考えられる．ここで,「新川」の認知度が高かった原因として,「新川」は地名として存在するため検索結果数が高くなったからであると考えられる．

次に，表 33 より，暗黙的観点「MLB の日本人選手」から抽出される対立語は「マック鈴木」,「高橋健」,「ダルビッシュ有」となった.「マック鈴木」や「ダルビッシュ有」はイチローと同程度の知名度を持つ語であり，対立語として適切であると考えられる．しかしながら,「高橋健」はイチローと同程度の知名度を持つ語であるとは言えず，対立語として不適切であると考えられる．認知度が高くなった原因として，お笑いコンビのキングオブコメディのメンバーである「高橋健一」と名前が重複しており，検索結果数に影響したと考えられる．

8.4 対立記事抽出の評価実験

本研究では，明示的観点に基づく対立記事，暗黙的観点に基づく対立記事の抽出を行い，評価実験を行った．

図 8: 明示的観点に基づく対立記事の評価の割合

8.4.1 明示的観点に基づく対立記事の評価実験

明示的観点に基づく対立記事が正しく抽出されているかを評価するため，**??**節と同様に表 21 に示すニュース記事 20 件に対してそれぞれ抽出された対立記事を 5 段階評価で実験を行った．この時，抽出された対立語それぞれに対し 8 名の被験者の評価の平均値をその対立語の評価値とする．

抽出された対立記事の評価値の割合を図 8 に示す．この結果より，評価値が 3.0 以上となった割合が 43 %となった．

対立記事の評価の高かった例としては,「VIP 会見にイチロー『ただただ恐縮している』」に関する記事に対して，対立記事が「『平成の怪物』松坂、悩んだ末の日本復帰…『世界一目指す』に魅力」に関する記事であった．評価の高かった理由としては,「会見」や「入団」,「球団」といった語が類似度計算に大きく影響し，閲覧記事と類似する記事が抽出されたと考えられる．

評価の低かった例としては,「ドイツ、１月の物価０・３％下落　５年４カ月ぶりにマイナス」に関する記事に対して，対立記事が「【ワイン】フランスがトップの供給国、次いでイタリア、チリ！」に関する記事となった．評価の低かった原因として，抽出した対立記事は「消費」という語のみが類似度計算に影響しており，抽出された対立記事以外の記事には閲覧記事と共通する単語がほぼ存在しなかったからであると考えられる．

8.4.2 トピック抽出に基づく暗黙的観点に基づく対立記事の評価実験

表 22 のニュース記事に対して観点毎に抽出した 8.3.2 節で抽出した対立語を用いて対立記事の抽出を行った．表 34 に記事 1 から抽出された暗黙的観点「火山」に基づく対立記事のタイト

ル，表 35 に記事 4 から抽出された暗黙的観点「打率」に基づく対立記事のタイトルを示す．

表 34: 記事 1 の暗黙的観点「火山」に基づく対立記事

ニュース記事	記事のタイトル
閲覧記事	阿蘇山で爆発的噴火 警戒レベルを２から３に引き上げ
対立記事	桜島の「無爆発」１７日まで１４４日間、 最長期間に並ぶ

表 35: 記事 4 の暗黙的観点「打率」に基づく対立記事

ニュース記事	記事のタイトル
閲覧記事	三塁打で達成は２人目９６年のモリター以来
対立記事	ヤクルト・山田哲人が２年連続「トリプルスリー」 ＆１００打点達成、史上初の快挙

表 34 に示すニュース記事は桜島が噴火してからの無爆発期間について述べた記事である．記事 1 は阿蘇山が爆発的噴火を起こしたニュース記事であるため，被害状況を比較することは出来ないが，阿蘇山の今後の状況と比較できるという点で対立記事として適切であると考えられる．また，表 35 に示すニュース記事は山田哲人が打率 3 割・30 本塁打・30 盗塁という「トリプルスリー」を 2 年連続で獲得したことについて述べた記事である．記事 4 はイチローが 3 千本安打達成に対するニュース記事であるため，抽出された対立記事では 3 千本安打の比較はできないと考えられる．そのため，対立記事として不適切であると考えられる．

8.4.3 概念構造に基づく暗黙的観点に基づく対立記事の評価実験

表 22 のニュース記事に対して観点毎に抽出した 8.3.3 節で抽出した対立語を用いて対立記事の抽出を行った．表 36 に記事 2 から抽出された暗黙的観点「日本の玩具メーカー」に基づく対立記事のタイトル，表 37 に記事 3 から抽出された暗黙的観点「日本の電気機器メーカー」に基づく対立記事のタイトル，表 38 に記事 4 から抽出された暗黙的観点「MLB の日本人選手」に基づく対立記事のタイトルを示す．

表 36: 記事 2 の暗黙的観点「日本の玩具メーカー」に基づく対立記事

ニュース記事	記事のタイトル
閲覧記事	任天堂、「Ｗｉｉ Ｕ」の生産を年内にも終了へ　次世代機に注力
対立記事	元阪神・金本知憲氏の１／６サイズのフィギュア　バンダイが限定生産

表**??**に示すニュース記事はバンダイが元プロ野球選手である金本知憲のフィギュア生産ついて述べた記事である．記事 2 は任天堂が Wii U を生産終了したニュース記事であるため，任天堂が Wii U を生産終了した裏でバンダイがフィギュア生産に着手しているという情報が得られ

表 37: 記事 3 の暗黙的観点「日本の電気機器メーカー」に基づく対立記事

ニュース記事	記事のタイトル
閲覧記事	日立が「レンズなし」のカメラ開発 フイルムでレンズ代用、薄く軽く低価格に
対立記事	日本品質レベルをクリアした光記録ディスクを輸入販売開始

表 38: 記事 4 の暗黙的観点「MLB の日本人選手」に基づく対立記事

ニュース記事	記事のタイトル
閲覧記事	三塁打で達成は２人目９６年のモリター以来
対立記事	ダルビッシュが２９日にメジャー復帰へ パイレーツ戦で先発見通し

るそのため対立記事として不切であると考えられる．また，表 37 に示すニュース記事は太陽誘電が生産する光記録ディスクを生産するための輸入開始に関するニュース記事である．記事 3 は日立製作所がレンズ無しのカメラ開発に関するニュース記事であるため，対立記事として不適切であると考えられる．表 38 に示すニュース記事はダルビッシュがメジャーリーグに復帰することに関する記事である．記事 4 はイチローが 3 千本安打達成に対するニュース記事であるため，イチローが 3 千本安打を達成している裏でダルビッシュが怪我をしていたという背景情報を得る事ができる．そのため対立記事として適切であると考えられる．

9　まとめと今後の課題

本研究では，ニュース記事の理解支援を目的として，観点に基づく対立記事の抽出手法を提案した．具体的には，まず閲覧記事から主題語を抽出する．次に閲覧記事から明示的観点を抽出し，主題語から暗黙的観点を抽出する．この時，暗黙的観点の抽出に対して２つの手法を提案した．一つは主題語を含むニュース記事群からトピック抽出を行うことで暗黙的観点を抽出する手法である．もう一つは主題語の上位概念から暗黙的観点を抽出する手法である．次に明示的観点，暗黙的観点それぞれから対立語を抽出し，対立語とそれぞれの観点を用いて対立記事を抽出した．また，ここで抽出した主題語の抽出，明示的観点の抽出，２つの暗黙的観点の抽出，対立語の抽出，明示的観点に基づく対立記事の抽出，２つの暗黙的観点それぞれに基づく対立記事の抽出に対して評価実験を行った．

また，今後の課題を以下に述べる．

- 明示的観点の抽出制度の向上
  本研究で抽出した明示的観点のうち，サ変活用名詞は明示的観点として適切でない場合が多いことが分かった．そのため，抽出の際にサ変活用名詞を除いた抽出を行う必要がある．

- 暗黙的観点の抽出制度の向上
  本研究では２つの手法によって暗黙的観点の抽出手法を提案した．評価実験により概念構造に基づく暗黙的観点の抽出手法がより有用であると分かったが，トピック抽出によっ

て抽出される暗黙的観点は抽出できない場合があることが分かった．そのため，それぞれの手法で抽出できる暗黙的観点を抽出する手法を提案する必要がある．

- ユーザインターフェースの作成
  本研究では，対立記事の抽出手法を提案したが，ニュース記事は Web 上で閲覧するため，対立記事も閲覧できるよう提示しなければならない．そのためユーザへの提示手法を考える必要がある．

謝辞

本研究を進めるにあたり，約 3 年半もの長きに渡り議論していただいた灘本先生に厚く御礼申し上げます．灘本先生のもとで研究をしていなければ，これまで研究を続けることが出来なかったと思います．そして，論文をまとめるにあたり，提出が遅れたにも関わらず査読してくださるだけでなく有益な御助言とご教授を賜りました甲南大学 小出 武先生，甲南大学 永田 亮先生に厚く御礼申し上げます．そして，研究に対して度々議論に参加してくださいました先輩方，後輩たちの皆にも厚く御礼申し上げます．最後に，大学院進学に向けて金銭的な支援だけでなく不自由なく生活をさせていただきました，父・大原正人，母・ 大原とも子に深く感謝致します．

研究業績

国内会議

- 大原 正章, 真下 遼, 灘本 明代, “Web ニュースの閲覧記事に対する対立記事抽出手法”, ARG 第 6 回 Web インテリジェンスとインタラクション研究会, 2015
- 大原 正章, 真下 遼, 灘本 明代, “Web ニュースからの観点抽出手法の提案”, 第 162 回データベースシステム研究会 (SIG-DBS), 2015

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, 2007.

[3] Keisuke Kiritoshi, Qiang Ma, “Named entity oriented related news ranking”, In Database and Expert Systems Applications, pp.82-96, 2014.

[4] 馬 強, 田中 克己, “補完情報の検索に基づくコンテンツ統合”, 情報処理学会研究報告データベースシステム（DBS）, 2004.72(2004-DBS-134),337 - 343,2004.

[5] 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.1-6,2014.

[6] 小林 透, 柴田 和樹, 中山 僚, “マルチコンテンツによる「ニュースの言葉」自動生成システムの研究”, マルチメディア、分散協調とモバイル (DICOMO2014) シンポジウム, pp.85-92,2014.

[7] 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, 2005.

[8] 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.

[9] 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, 2011.

[10] David M Blei, Andrew Y Ng, Michael I Jordan, “Latent dirichlet allocation”, the Journal of machine Learning research, Vol. 3, pp.993-1022, 2003.

[11] 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.

[12] 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.