3M-01

Webニュースの対立記事の抽出手法

大原　正章†　　灘本　明代†

甲南大学知能情報学部†

1　はじめに

インターネットの普及によりWeb上のニュースサイトを利用するユーザが増加している．ニュースの情報を得る手段として，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 ニュースを読む際，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，ある企業の赤字に関する記事を閲覧した際に，その企業の事は知っていてもその業界についての知識が無い場合，閲覧記事がどのくらい重大であるのかを理解することは容易ではない．このような場合，その企業のライバル企業の収支に関する情報と比較することにより，閲覧中の記事の重要性を把握することが可能である．

一方，現在のWebニュースには，その記事に対する関連記事のリンクが存在する．しかしながら，その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する記事である場合がほとんどである．このような関連記事は，経過情報を記載したニュースやニュースの主題が同じというだけで内容が全く関連しないニュースであることが多い．そのため，関連ニュースを閲覧した場合においても，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（ライバル）関係にある記事を提示することにより，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，本研究では，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．

具体的には，ユーザが閲覧している記事から主題となる単語と記事の観点（記事アスペクト）を抽出し,その主題と対立する単語に対して閲覧記事と同じ記事アスペクトを持つ記事を抽出して提示する手法を提案する．

2　関連研究

本研究では閲覧しているニュース記事に対する対立記事を発見することを目的としている．

田中ら[1]は，ニュース記事から人物や組織，場所等の重要な特徴語をエンティティとして取得し，Wikipedia のデータを背景知識として取得し，補完する手法について提案している．また，エンティティのランキングを行う際にはTextRank に類似する手法を用いている．本研究ではエンティティではなくアスペクトを用いており，そのランキングの手法が異なる．しかしながら，閲覧記事の理解を支援するという目的が一致している．

北山ら[2]は，ニュースメディア毎に異なる記事の位置付けに着目し，コンテンツ構成要素について順序特性を付与した質問を生成している．生成された質問から，異なるメディアで閲覧記事とは位置付けの違った同一の出来事に関する記事を抽出して比較記事として提示する手法を提案している．本研究では抽出する記事の主題が閲覧記事の主題とは一致しない点で異なるが，関連した記事を抽出する点や記事の構造に着目している点で類似している．

真下ら[3]は，漫才台本として対立ボケの生成手法を提案している． 対立ボケとは，ある語に関して対照的な関係にある単語を用いて行うボケである．対立する単語の取得を行う際にWikipedia の階層構造を利用しており，本研究でもこの手法を参考にした．

3　提案手法

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)閲覧記事から記事アスペクトの候補を抽出する．
(3)主題語に対してライバルとなり得る対立語の候補を抽出する．
(4) (2)で抽出した記事アスペクトの候補と(3)で抽出した対立語の候補との共起関係からスコアリングを行う．
(5) (4)の結果，最も高いスコアを持った対立語の候補と記事アスペクトの候補のペアを対立語と記事アスペクトとする．
(6)対立語を主題とし，ここで決定した記事アスペクトを持つニュース記事を取得し，対立記

Extracting of Rival News Article from Web News
Masaaki Ohara†
Akiyo Nadamoto†
Dept. of Intelligence and Informatics Factly of Intelligence and Informatics††

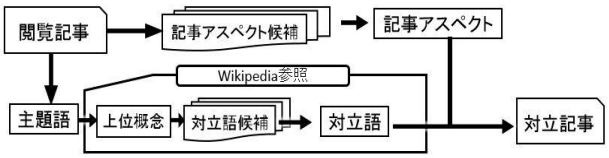

図1 提案手法の概要

事としてユーザに提示する．

以下，各々の手順について詳しく述べる．

3.1 主題語の抽出

ユーザが閲覧している記事から主題語と記事アスペクト候補を取得する．ニュースの主題語は，そのニュースに頻出している単語で且つ固有名詞である場合が多いと考え，固有名詞の単語の出現頻度の高い物とする．単語の出現頻度を求める際，ニュース記事のタイトルでは最も重要な事物を記載しており，本文の1段落目では記事の内容が簡潔にまとめられ，本文の2段落目以降では付随する内容が記載されている点に着目し，タイトルと本文の1段落目，本文の2段落目以降で各々出現回数に重みを付与して主題語の抽出を行う．

3.2 記事アスペクト候補の抽出

記事アスペクトは，主題語に対するその記事の観点であると考える．この場合，一つの記事に対して複数の観点がある場合がある．そこで，記事アスペクト候補としてすべての観点に対して抽出を行う．この観点は記事の特徴となる単語であることが好ましい．そのため記事アスペクトとなる単語も記事内では主題語と同じ程度に重要であると考え，固有名詞以外の名詞の出現頻度を算出して閾値以上の全ての名詞を記事アスペクト候補とする．ここで，3.1 と同様に単語の出現位置を考慮した重みの付与を行う．

3.3 対立語候補の抽出

我々の提案する対立語は，「阪神タイガース」と「巨人」，「野球」と「サッカー」といったように，２つの語同士が互いに対照的な関係性にある語のことを指す．２つの語の関係性に着目すると「阪神タイガース」と「巨人」は共にプロ野球チーム，「野球」と「サッカー」は共に球技のように語同士が共通の上位概念を持っていることがわかる．また，「野球」の同位語として「サッカー」と「フットサル」が挙げられるが，「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」の方が対立語として適切であると考えられる．これらを踏まえて本研究では対立語を，キーワードの同位語であり，同程度の認知度を持つ語と定義し，対立語の候補を抽出する．まずは主題語の上位概念を取得するために，Wikipediaの階層構造をコーパスとして用いる．ここで例えば，東京では「日本の都市」や「就航地」，「曲」など11語の上位概念が取得できる．取得した全ての上位概念に対し，各々の下位概念つまりは主題語に対して兄弟概念となる単語を対立語予備候補とする．次に この対立語予備候補と主題語の認知度は検索結果数を用いる．そこで対立語予備候補と主題語の検索結果数がある閾値以内に類似している対立語予備候補を対立語候補とする．

3.4 記事アスペクトと対立語の決定抽出

対立語記事を求めるために対立語候補から，閲覧記事の記事アスペクトと関係の深い語を対立語とする．本論文では，対立語候補と記事アスペクトがよく共起している場合，2 つの語の関係が深いと見なす．具体的には，3.2 で取得した記事アスペクト候補と 3.3 で取得した対立語候補のシンプソン係数を求め，それを２つの語の関係のスコアとする．このスコアが最も高い組み合わせを対立語と記事アスペクトと決定する．

3.5 対立記事の抽出

対立記事の主題語は対立語であることが望ましいため，対立語と記事アスペクトを含む記事から主題を取得し対立語と一致する記事を対立記事とする．

4 まとめ

ニュース記事の重要性の理解を支援することを目的とし，対立する記事を提示するための手法を提案した．具体的には，まず閲覧記事から主題語と記事アスペクト候補を取得した．次に Wikipedia の階層構造を用いて主題語に関する対立語の候補を抽出し，記事アスペクト候補と対立語候補の共起関係から，記事アスペクトと対立語を決定する．そして取得した対立語と記事アスペクトから対立記事の提示を行った．

今後の課題は，評価実験を行い，提案手法の有用性を示す．

参考文献

[1]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，情報処理学会研究報告 Vol2014-DBS-159 No.17，6pages，2014

[2] 北山 大輔，角谷 和俊，コンテンツ構成要素の順序特性に基づく比較ニュース検索方式，信学技報 IEICE Technical ReportDE2007-68 (2007・7)，pp.277－282

[3] 真下 遼，灘本 明代，対立語抽出に基づく Web ニュースからの漫才ロボット台本自動生成手法の提案，DEIM Forum 2014 C2-4，8pages

[4] 産経ニュース，http://www.sankei.com/