ユーザの興味を考慮したWebニュース理解支援のための例え表現生成手法

真下　遼†　灘本　明代†

† 甲南大学 知能情報学部　〒 658–8501 兵庫県神戸市東灘区岡本 8-9-1
E-mail: †{m1424008,nadamoto}@konan-u.ac.jp

あらまし　本論文では，ユーザのニュースに対する興味喚起および内容理解の向上を目的に Web ニュースのタイトルをユーザの興味に合わせた文で例える，例え表現生成手法を提案する．例え表現生成では，Web ニュースのタイトル中に含まれる単語の価値と文構造に着目し，価値が近似するユーザの興味カテゴリに属した単語を抽出し用いることで Web ニュースの本質的情報を保持しながらユーザにとって理解しやすい例え表現を自動生成する．また，生成した例え表現に関して評価実験を行い提案手法の有効性を確認する．

キーワード　ニュース，Web，PageRank，Wikipedia

1. はじめに

近年インターネットの利用が普及すると共に，総務省のインターネットの利用目的に関する調査 (注1) では，ニュースサイトの利用が年代を問わず高い割合を占めている．ニュースサイトでは，Web ニュースの配信によりユーザは多種多様な分野の情報をいつでも手軽に取得することが可能である．Web ニュースは不定期に随時配信され続けるため，TV のニュース番組や新聞をも超える高い速報性を維持している．そのため最新の情報を Web ニュースから抽出することが増加し，現在 Web ニュースは主要な情報源の一つになっていると考えられる．ニュース記事が手軽に取得できる一方で，多種多様な情報をユーザが的確に把握し内容を理解するのは困難であると考えられる．例えば,「羽生結弦がフィギュアスケート NHK 杯で 3 連覇」という内容の Web ニュース記事の場合，フィギュアスケートに関しての知識や関心が少ない閲覧ユーザはこの記事の重大性やニュースになった理由を理解することは難しいと考えられる．

これまでに閲覧記事から適切な関連記事を推薦する研究 [1] や閲覧記事の自動要約および要点の抽出を試みる研究 [2] が多数行われている．これらの研究は，いずれもニュースの閲覧ユーザの理解支援を主な目的としたものであるが，関連記事の閲覧はユーザにさら多くの文書を読む負担を強いる必要がある．記事の要約においては，ユーザがニュースのテーマに関する予備知識がない場合，閲覧ユーザはタイトルを読んだ時点で閲覧記事への関心が薄れてしまい結果的に閲覧記事ジャンルの偏りが生じる可能性がある．そこで我々は，ニュース記事への興味喚起および理解支援の手法としてユーザ自身の興味を考慮した例え表現に着目した．例え表現とは，広義には「ある物事を別の似ている物事で表現すること」であり，これを我々は Web ニュース記事に適用することを考えた．先の「羽生結弦がフィギュアスケート NHK 杯で 3 連覇」のニュース記事を例にすると，野球に関する知識や興味のあるユーザには「上原浩治が日本選手権シリーズで優勝」と例えて提示することで記事の重要性がより直感的に理解できると考えた．そこで本論文では，あるニュースのテーマに関して知識のない閲覧ユーザに対する理解支援および興味喚起を目的とし，ニュースを閲覧ユーザの知識または興味のあるテーマで例える，例え表現を生成し提示する手法を提案する．

本論文では例え表現を行う対象としてニュースのタイトルに着目し，例え表現を生成する手法を提案する．これは，ニュースのタイトルがニュース記事の閲覧時にユーザが最初に注目すると思われる箇所である点や，ニュース記事の性質上タイトル中に重要な語が出現するといった点を考慮したためである．例え表現を自動生成するに当たり，タイトルが「羽生結弦がフィギュアスケート NHK 杯で 3 連覇」といったように「“誰/何 ’’ が “どこ/何 ’’ で “どうした ’’ 」の「S（主語）が O（目的語）に V（述語）した」となっていることに着目する．先の野球での例え表現では，S（主語）の羽生結弦を上原浩治で表現し，同様に O（目的語）のフィギュアスケート NHK 杯を日本選手権シリーズで表現している．この時，上原浩治の部分および日本選手権シリーズの野球の部分はユーザの興味のあるテーマに関連していることがわかる．また，同じ野球でも「上原浩治が “日本選手権シリーズ ’’ で優勝」で例えるのと「上原浩治が “神奈川県野球交流戦 ’’ で優勝」で例えるのとでは勝利の価値に開きが生じ，記事内容の理解を妨げる恐れがある．この場合では，元の記事のフィギュアスケート NHK 杯と大会の権威的価値がある程度近似している日本選手権シリーズで例える方が比較的相応しいと考えられる．そのため，例え表現の生成において，例える対象であるものの価値が各々の分野においてある程度同等である必要があると考える．以上を踏まえて，本論文では,「ユーザの興味」と「ものの価値」を考慮した Web ニュース記事のタイトルから例え表現生成手法を提案する．我々の提案する例え表現生成は，ユーザがシステムに自分の興味のある

(注1)：http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h27/html/nc372110.html

テーマを入力すると，ユーザが閲覧している記事のタイトルとユーザが入力したテーマをもとにしてシステムが例え表現を生成し提示する．

ユーザの興味とものの価値を考慮した Web ニュース記事からの例え表現自動生成では以下の点で有用であると考えられる．

- Web ニュース記事内容の理解支援．

Web ニュース記事内の情報とは異なるユーザの興味に関連した情報を提供することで，ユーザは新たな観点からニュース内容の理解を行うことが可能である．

- 情報視野の拡大．

ユーザの固有の知識や興味に合わせた例え表現により元記事への興味喚起を促すことが可能である．

- 異なるカテゴリ間の新たな関係構築．

例え表現では本来関係性を見出し難い語の組み合わせに対して，価値の近似やユーザの興味という観点から新たな関係を見出すことが可能である．

これまでにも我々は例え表現の自動生成手法 [3] を提案してきたが，本論文ではシステムの更なる改良行うとともに，実際にシステムを実装してユーザ実験を行い提案手法の有用性について報告する．

以下，第 2 章では関連研究について述べ，第 3 章では例え表現自動生成について述べ，4 章で評価実験およびその考察について述べ，5 章でまとめと今後の課題について述べる．

2. 関 連 研 究

本論文のようにニュース情報の理解支援を目的とした研究は多数行われている．平田ら [4] はニュース記事内に記述されているイベント情報を抽出しユーザの興味に合わせた系列を構成して提示することでニュース記事の閲覧を支援するシステムを提案している．張ら [5] はニュース記事の印象に着目して，ニュースの印象をユーザに提示することでニュースサイトの報道傾向を視覚的に比較できるようにしている．石井ら [6] はニュースによる事象間の因果関係を SVO の文法構造に着目してネットワーク構造で表現す手法を提案している．Souneil [7] らは，アスペクトに着目して，NewsCube と呼ばれるニュース記事閲覧サービスを開発し，情報操作の緩和を目指している．北山ら [8] は，ユーザが閲覧記事に関連す る記事をアーカイブから的確に抽出するための検索方式を提案している．ニュースに限らず Web 上の情報の理解支援を目的とした研究も多数ある．Oyama ら [9] は Web ページの文章構造をからある語を詳細に説明する詳細語を発見する手法を提案している．西原ら [10] は，未知語の理解支援として Web ページの検索結果をクラスタリングすることで，未知語の難易度とそのための学習順序を検出する手法を提案している．佃ら [11] はある語に関してより上位語らしい語と同意語らしい語を抽出しユーザに提示することで理解支援を行っている．本論文では例え表現に着目し，閲覧記事のタイトルをユーザの興味に合わせた別の情報で表現することで，閲覧記事への興味換気からの理解支援を目指している．また，ユーザの閲覧の負担を抑えたものとなっている．

例え表現生成と類似する研究として，言い換え表現の自動生成に関する研究 [12] 等が様々あるが，言い換え表現とは，ある言語表現の意味を保ったまま表現を変換するのに対して，本論文で提案する例え表現では意味ではなく価値観を保つことに加えユーザの興味を重要視している点で大きく異なる．

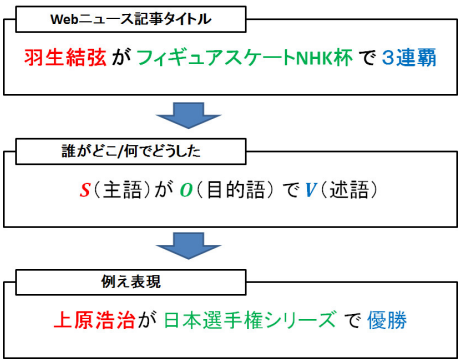

図 1　例え表現生成イメージ

価値に着目した研究として木虎ら [13] は，ユーザの Web アクセス履歴からユーザ固有の価値観を分類を行っている，奥ら [14] は，価値判断基準のモデルをユーザの嗜好及びその時のユーザの状況を表すユーザコンテキストに着目して独自に定義している．本論文では，ユーザが未知としている単語も対象にして価値を推定し，その価値が近似している単語を最終的提示するため，ユーザ固有の価値観ではなく客観的・社会的価値観を推定することを目的としている点が異なる．

3. 例え表現の自動生成

3.1 例え表現の定義

例え表現の生成のために例え表現をある程度形式的に捉える必要がある．図 1 に本論文における例え表現生成のイメージ図を示す．本論文では，Web ニュースのタイトルが「羽生結弦がフィギュアスケート NHK 杯で 3 連覇」のように，「“誰/何 ’’ が “どこ/何 ’’ で “どうした ’’ 」の「S（主語）が O（目的語）に V（述語）した」となっていることに着目する．ここで，変換元となるユーザが閲覧中の記事タイトルの各「S（主語）」，「O（目的語）」，「V（述語）」に出現する単語をそれぞれ S_0，O_0，V_0 とすると，「羽生結弦がフィギュアスケート NHK 杯に 3 連覇」は S_0 が「羽生結弦」，O_0 が「フィギュアスケート NHK 杯」，V_0 が「3 連覇」と表せる．さらに，例え表現中の各「S（主語）」，「O（目的語）」，「V（述語）」をそれぞれを S_1，O_1，V_1 とすると，先の例を野球で例えた場合の「上原浩治が日本選手権シリーズで優勝」は S_1 が「上原浩治」であり，O_1 は「日

表 1　例え表現に用いる語とその定義条件

例え表現語	定義条件
S_1	ユーザの知識や興味のカテゴリに属する語
	閲覧記事タイトルの S_0 と価値が近似している語
O_1	ユーザの知識や興味のカテゴリに属する語
	閲覧記事タイトルの O_0 と価値が近似している語
V_1	S_1 と O_1 に対して文脈的繋がりを持つことができる語
	閲覧記事タイトルの V_0 と意味が類似している語

本選手権シリーズ」となる．つまりは,「羽生結弦→上原浩治」,「フィギュアスケート NHK 杯→日本選手権シリーズ」に変換している．変換した単語同士は，各々単語の示す価値が類似していることがわかる．即ち，S_0 と S_1，O_0 と O_1 の価値が類似している方が，元の記事の重要性がよりわかりやすいと考える．なお本論文では，単語の価値を「ものの価値」と呼ぶ．

文の結論を示す役割にあたる述語の V_0 と V_1 は意味が類似している語が望ましいと考えられる．例えば,「上原浩治が日本選手権シリーズに苦戦」と例えた場合元の記事の示す内容を大きく損なうものとなる．例え表現生成のために必要な単語 S_1，O_1，V_1 の定義条件をまとめたものを表 1 に示す．本論文ではニュースのタイトルを構成する単語 S_0，O_0，V_0 に着目すると共に，例え表現に用いる単語 S_1，O_1，V_1 をものの価値と意味の類似を考慮して抽出し，対応する語同士を置換することで例え表現を生成する．また，本論文で生成する例え表現では，生成する表現の自由性を求める目的で生成するため，例え表現の内容が事実であるかどうかは問わないものとする．

3.2 価値の推定

本論文では，ものの価値を考慮し，例え表現を自動生成する．そこで，まずものの価値を推定により定量化し，定量化した価値の値を比較することで価値が類似している語を抽出する．ものの価値の定量化を行う場合，経済的側面での数値評価を利用することが考えられるが，ものの価値は対象によって様々な指標や視点により決まるため定量化は困難である．また，本論文で提案する例え表現においては経済的側面での評価が行われないものも評価対象とするため，経済的側面での価値の評価は行わない．本論文では，ものの価値を判断する指標として以下の 2 つの仮説を立てる．

- 価値あるものには価値あるものが関わる．
- 価値が時間的に持続しているものはより価値がある

これらの 2 つの仮説を組み合わせてものの価値を定量化する．

価値あるものには価値あるものが関わる

例えば，権威ある賞の歴代受賞者には多くの著名人が受賞し名誉を得ている．同時に各著名人達が受賞者として関わることで賞自体もその権威を高めていると考えられる．これは賞に限ったものではなく，大会や人間関係等様々なものにおいても同様の関係が見られる．例え表現におけるものの価値を定量化するために汎用性の高い指標の一つとしてこの仮説は有用であると考えた．この仮説に基づく指標の定量化として PageRank アルゴリズム [15] を用いる．PageRank アルゴリズムは Web のハイパーリンク構造を用いて Web ページを順位付けするアルゴリズムであるが，PageRank アルゴリズムの根本的な考え方として，多くの良質なページからリンクされているページはやはり良質なページであるという考えがある．この考え方に基づき,「価値あるものには価値あるものが関わる」として PageRank アルゴリズムを用いる．本論文では，価値を定量化する対象の語 t_i の Wikipedia 記事 W_i とその記事間のリンク関係を用いて以下の式 (1) により t_i の価値 $PR(t_i)$ を算出する．

$$PR(t_i) = (1-d) + d\sum_{j=1}^{n} \frac{PR(P_j)}{C(P_j)} \quad (1)$$

ここで，n は記事 W_i へリンクしている記事の総数，$C(P_j)$ が記事 W_i と記事 P_j 以外の記事へのリンクする記事の総数であり，$PR(P_j)$ が記事 W_i にリンクしている j 番目の記事の PageRank を表す．また，d はダンピングファクターで，通常用いられるようにここでは 0.85 を設定する．なお，記事データおよびそのリンク構造は日本の Wikipedia 情報ダウンロードページ (注2) から 11 月 1 日に抽出したものを用いる．本論文では，$PR(t_i)$ の値が高い語 t_i ほど，ものの価値が高い語とみなす．

価値が時間的に持続しているものはより価値がある

流行や風化といった言葉が表すように，物事の価値は日々時間の影響を受けながら変化していき一定ではないと考えられる．本論文での例え表現がニュースを対象としてた理解支援の目的があることを考えると，流行や風化といった突発的に価値を有しているものよりも，継続的に価値を有しているものをユーザに提示する方が望ましいと考えられる．そこで本論文では，時間的な情報も加味して価値の評価を行うことを考える．具体的には，Wikipedia 記事の閲覧回数がその記事の社会的関心を表しているものと仮定して，価値を評価する対象の語 t_i の Wikipedia 記事の閲覧回数を利用する．対象の記事の現在から過去 5 年間に及ぶ月毎の閲覧回数を抽出し，その中央値を価値評価の評価指標とする．中央値の値が高い語ほど，価値が高い語と見なす．

3.3 ものの価値の決定

上記 2 つの仮説の元，単語 t_i の価値 $Val(t_i)$ を以下の式にて決定する ．

$$Val(t_i) = PR(t_i) + \log TD(t_i) \quad (2)$$

ここで，$PR(t_i)$ は，1 つ目の仮説に基づいて Wikipedia 記事のリンク関係から PageRank アルゴリズムにより抽出する単語 t_i の関係性の価値を意味する．$TD(t_i)$ は，2 つ目の仮説に基づいて，月毎の閲覧回数により抽出する単語 t_i の時間性を考慮した価値を意味する．最終的に式 2 により単語の価値を算出し定量化したものがその単語のものの価値とする．$Val(t_i)$ の値が高い語 t_i ほど，ものの価値が高い語とみなす．

3.4 例え表現自動生成手順

本論文で提案する例え表現の生成の全体の流れを図 2 に示す．また，例え表現の生成の流れの概要を以下に示す．

（1） ユーザは自身の興味 T を入力する．同時にシステムは閲覧中の Web ニュース記事のタイトルを抽出する．

（2） 抽出した Web ニュース記事タイトル中から S_0，O_0，V_0 をそれぞれ抽出する．

（3） 例え表現の主語 S_1 を抽出する．

（4） 例え表現の目的語 O_1 を抽出する．

(注2)：http://download.wikimedia.org/jawiki/latest/

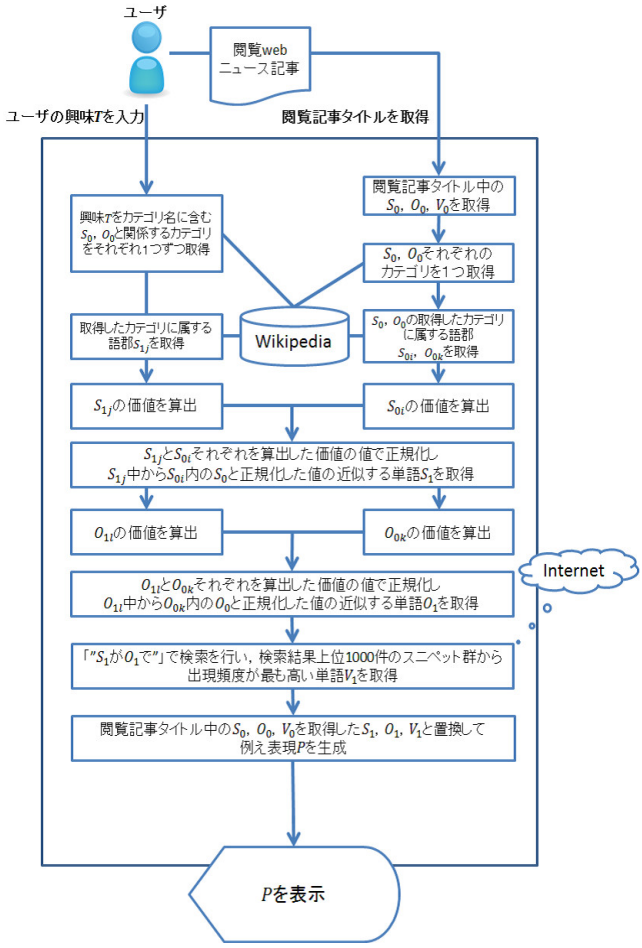

図 2　例え表現自動生成システムフロー

（5） 例え表現の述語 V_1 を抽出する．

（6） 抽出した S_1，O_1，V_1 を用いて例え表現 P を生成しユーザに提示する．

我々の提案する例え表現生成システムでユーザが行うことはキーワードの入力のみである．ユーザが入力するクエリ T は，例えば「野球」や「ゲーム」といったカテゴリを表すような 1 単語を想定しており，ここで入力した T を以降ユーザの興味とする．次に，例え表現生成のために置換するべき語となる S_0，O_0，V_0 を閲覧中の Web ニュース記事のタイトル中から係り受け解析を行い抽出する．ここで，係り受け解析には CaboCha(注3) を用いた．S_0，O_0，V_0 の抽出を行った後，例え表現生成に必要となる S_1，O_1，V_1 を順に抽出する．最後に，抽出した S_1，O_1，V_1 を用いて閲覧中の Web ニュース記事のタイトルの例え表現を生成しユーザに提示する．以降，S_1，O_1，V_1 の抽出手法について詳しく述べる．

例え表現の主語 S_1 の抽出

例え表現の主語となる S_1 は，表 1 の定義条件に示すように「ユーザの興味」と「ものの価値」を考慮して抽出する．ここで，ユーザの興味については事前にユーザからの入力により T が与えられているため，この T に属する語を S_1 の候補とすることで解決する．

価値の考慮においては，S_0 と S_1 の価値が互いに近似している関係が望ましいと考えられる．ただし，先に述べたように S_1 は同時にユーザの興味 T に属する語である必要があり，ユーザの興味によっては S_0 と価値が近似している S_1 が必ずしも存在しない可能性がある．そこで本論文では，S_0 と S_1 のそれぞれで同じカテゴリに属する語群 S_{0i} $(i = 1, \ldots, n)$ と S_{1j} $(j = 1, \ldots, m)$ を抽出して二つのグループを生成する．そして，各グループ同士で価値の値で正規化を行い，正規化した値をグループ間で比較することにより S_0 に価値が近似する S_1 を一意に決定する．以下に S_1 の抽出の流れを示す．

（1） S_0 の属するカテゴリを一つ抽出し，そのカテゴリに属する語群 S_{0i} を抽出する．

（2） ユーザが入力した興味 T をカテゴリ名に含むカテゴリを一つ抽出し，そのカテゴリに属する語群 S_{1j} を抽出する．

（3） 抽出した S_{0i} と S_{1j} の全てのものの価値を推定し算出する．

（4） S_{0i} と S_{1j} をそれぞれ別々に算出した価値で正規化する．

（5） 価値で正規化したそれぞれの値を比較し，S_{0i} 中の S_0 と近似する価値を持つ語を S_{1j} 中から S_1 として抽出する．

例として閲覧中の Web ニュース記事のタイトルを「羽生結弦がフィギュアスケート NHK 杯で 3 連覇」，ユーザの興味の入力を「野球」とした場合，S_0 に当たる語は羽生結弦である．

まず，S_0 の羽生結弦が属するカテゴリを一つ抽出する．本論文では，カテゴリを Wikipedia のデータ構造を用いて抽出する．羽生結弦の Wikipedia 記事が属するカテゴリとして Wikipedia のデータ構造からは，「日本の男子シングルスケーティング選手」，「「仙台市出身の人物」，「存命人物」等の計 10 個のカテゴリが抽出できる．このように羽生結弦の一つの Wikipedia 記事に対して複数のカテゴリが得られるが，最終的に同一カテゴリに含まれる語群で正規化することを考えると羽生結弦を最も的確に表現しているカテゴリを抽出することが望ましいと考えられる．ここで，Wikipedia 記事の最初の 1 文がそのタイトルの概要を顕著に表している [16] ことに着目する．羽生結弦の Wikipedia 記事の最初の 1 文は，「宮城県仙台市泉区出身の日本人フィギュアスケート選手」となっている．ここでさらに，Wikipedia 記事の最初の 1 文の最後に出現する一連の名詞である「日本人フィギュアスケート選手」が羽生結弦のカテゴリとして相応しいと仮定する．即ち，「日本人フィギュアスケート選手」が羽生結弦を最も的確に表現しているカテゴリであると考え，先の Wikipedia のデータ構造から抽出した羽生結弦が属する 10 個のカテゴリには「日本人フィギュアスケート選手」というカテゴリは存在しない．そこで，「日本人フィギュアスケート選手」と Wikipedia のデータ構造から抽出した羽生結弦が属する 10 個のカテゴリ名との文によるコサイン類似度により類似性を測り，最も類似度が高いカテゴリを羽生結弦を最も的確に表現していると思われるカテゴリとする．今回の場合，羽生

(注3)：http://chasen.org/ taku/software/cabocha/

表 2　オリンピックフィギュアスケート日本代表選手に属する語 S_{0i}（左）と
オリンピック野球日本代表選手に属する語 S_{1j}（右）の価値とその正規化

オリンピックフィギュアスケート日本代表選手	$Val(t_i)$	正規化 $(Val(t_i))$	オリンピック野球日本代表選手	$Val(t_i)$	正規化 $(Val(t_i))$
荒川静香	0.00449	1.0	野茂英雄	0.01323	1.0
浅田真央	0.00439	0.9769	松坂大輔	0.01204	0.9072
安藤美姫	0.00360	0.8023	田中将大	0.01132	0.8531
井上怜奈	0.00227	0.5060	森野将彦	0.00748	0.5618
村主章枝	0.00225	0.5025	古田克也	0.00706	0.5318
羽生結弦	0.00214	0.4778	上原浩治	0.00621	0.4675
鈴木明子	0.00174	0.3889	宮本和知	0.00614	0.4618
本田武史	0.00163	0.3628	松中信彦	0.00602	0.4549
八木沼純子	0.00162	0.3616	ダルビッシュ有	0.00570	0.4358
小塚崇彦	0.00148	0.3302	小笠原道大	0.00511	0.3872

結弦のカテゴリとしてコサイン類似度の最も高かった「オリンピックフィギュアスケート日本代表選手」を抽出する．なお，コサイン類似によってもカテゴリの優劣が付かない場合は，そのカテゴリに属している語が最も少ないカテゴリが羽生結弦と最も密に繋がっているカテゴリと仮定して抽出する．カテゴリを抽出した後，そのカテゴリに属する語群を抽出する．「オリンピックフィギュアスケート日本代表選手」のカテゴリに属する語には，「浅田真央」，「安藤美姫」，「小塚崇彦」等計 60 語が得られ，これらが S_{0i} となる．

次に，ユーザ入力した興味 T を基にして S_{1i} の抽出を行う．先の例では，興味 T は「野球」であったが，「野球」をカテゴリ名に含むカテゴリは Wikipedia のデータ構造では，「日本の野球大会」，「日本の野球選手」，「プロ野球チームの経営者」等の計 224 個のカテゴリが得られる．最終的にこれらのカテゴリに含まれる語群の中から S_1 を抽出することを考えると，S_0 と一部のカテゴリ属性が一致しているカテゴリを抽出することで，より例え表現に的確な S_1 を抽出することができると考えられる．そこで，羽生結弦の最も的確に表現しているカテゴリとして「オリンピックフィギュアスケート日本代表選手」を先に得たのと同様に「野球」をカテゴリ名に含むカテゴリとの文によるコサイン類似度により類似性を測り，最も類似度が高いカテゴリを抽出する．ここでは，「オリンピック野球日本代表選手」を一意に抽出し，そのカテゴリに属する語として「野茂英雄」，「松坂大輔」，「上原浩治」等計 54 語が得られ，これらが S_{1j} となる．

次に，S_{0i}，S_{1j} の語の全てに対して式 2 により価値を算出する．さらに S_{0i}，S_{1j} 毎にそれぞれ最小値が 0，最大値を 1 として 0-1 の範囲で値を正規化する．表 2 に S_{0i} の「オリンピックフィギュアスケート日本代表選手」と S_1 の「オリンピック野球日本代表選手」に属する語とその価値の値および正規化した値の一部をそれぞれ表記する．表 2 から S_{0i} 中の S_0 の「羽生結弦」の $Val(t_i)$ を正規化した値 0.4778 に最も値が近似する S_{1j} の語は 0.4675 の「上原浩治」であり，結果的に「上原浩治」を S_1 として抽出する．

例え表現の目的語 O_1 の抽出

例え表現の目的語となる O_1 も S_1 と同様の手法により「ユーザの興味」と「ものの価値」を考慮して抽出する．先の Web ニュース記事タイトル「羽生結弦がフィギュアスケート NHK 杯で 3 連覇」での O_0 は「フィギュアスケート NHK 杯」となる．「フィギュアスケート NHK 杯」のカテゴリの抽出および O_{0k} の抽出は S_0 から S_{0i} を抽出したのと同様の手法で行われ，結果的にカテゴリは「NHK 杯」を抽出し，O_{0k} には「NHK 杯テレビ将棋トーナメント」，「全日本相撲選手権大会」等の計 30 個の語が抽出される．

次に O_{1l} の抽出を行う．これも，S_{1j} の抽出と同様の手法でユーザの興味 T の「野球」と「NHK 杯」のカテゴリ情報を考慮して，コサイン類似度によりカテゴリを抽出する．しかしながら，「「野球」を含むカテゴリでは「NHK 杯」とのコサイン類似度が全て 0 となりカテゴリを一意に抽出することができない．そこで「NHK 杯」と意味的に類似する語を新たに取得し類似を測る指標を拡張する．具体的には，「NHK 杯」の「杯」のようにカテゴリの末尾の語がそのカテゴリの意味合いを強く担っていると仮定して，「杯」と意味的に類似する語を新たに取得する．本論文では，意味的に類似する語の取得に word2vec(注4) を用いる．word2vec とは，単語の意味をベクトル表現したものを獲得する手法である．なお本論文では，単語の意味ベクトルを Wikipedia の全文章コーパスを用いて学習する．word2vec により「杯」と単語間距離が近い語を上位 10 語取得する．ここでは，「杯」と単語間距離が近い語として「カップ」，「選手権」，「大会」等を取得する．この内「大会」とのコサイン類似度により「日本の野球大会」を「野球」を含むカテゴリから抽出する．さらに，「日本の野球大会」に属する語である「クライマックスシリーズ」，「日本選手権シリーズ」等の計計 32 個の語を O_{1l} として抽出する．なお，word2vec を用いてもカテゴリが抽出できない場合は，O_0 をそのまま O_1 として用いる．

その後，O_{0k}，O_{1l} の全ての語に対して式 2 により価値を算出し O_{0k}，O_{1l} 毎にそれぞれ正規化する．結果的に O_0 の「フィギュアスケート NHK 杯」に対して正規化の値が最も近似する語である「日本選手権シリーズ」を O_1 として抽出する．

例え表現の述語 V_1 の抽出

例え表現の述語 V_1 の抽出は，表 1 に示すように意味的ベクト

(注4)：word2vec, https://code.google.com/p/word2vec/

ルと文脈的つながりを考慮して行う．具体的には「“S_1 が O_1 で ’’ 」をクエリとして検索を行い，検索結果上位 1000 件のスニペットの中で「“S_1 が O_1 で ’’ 」に続く名詞あるいは動詞を V_1 として抽出する．先の例の場合，S_1 が「上原浩治」，O_1 が「日本選手権シリーズ」となっているため「“上原浩治が日本選手権シリーズで ’’ 」をクエリとして検索を行う．しかしながら，独自に抽出した S_1，O_1 でのクエリに対して続く名詞や動詞を抽出できることは希であり今回の「“上原浩治が日本選手権シリーズで ’’ 」の場合においても抽出することはできない．

そこで本論文では，V_1 の候補 V_{1l} を先の O_1 のカテゴリの取得に利用したのと同様に word2vec を用いて取得する．word2vec により V_0 と単語間距離が近い語上位 10 語を取得し V_{1l} とする．この時，V_{1l} に V_0 も含まれる．次に，V_{1l} の中から文脈的繋がりを考慮して，一意に V_1 を抽出する．具体的には，O_1 と V_{1l} の検索結果件数の共起頻度を Simpson 係数により測り，O_1 と最も共起する V_{1l} を V_1 として抽出する．表 4 に V_0 の「3 連覇」に対して word2vec により取得した V_{1l} を示す．また，V_{1l} と O_1 の「日本選手権シリーズ」との Simpson 係数も合わせて示す．表 4 より「日本選手権シリーズ」と最も共起度の高い「優勝」を V_1 として抽出する．

以上により S_1 が「上原浩治」，O_1 が「日本選手権シリーズ」，V_1 が「優勝」となり，これらを閲覧ニュース記事タイトル中の S_0，O_0，V_0 と置換することで例え表現 P を生成する．結果的に,「羽生結弦がフィギュアスケート NHK 杯で 3 連覇」に対する例え表現 P は「上原浩治が日本選手権シリーズで優勝」となる．

表 3 「3 連覇」の類似語と Simpson 係数

類似語	Simpson 係数
3 連覇	0.07
制覇	0.12
勝利	0.30
優勝	0.62
決勝	0.18
ワンツーフィニッシュ	0.02
チャンピオン	0.17
雪辱	0.03
快勝	0.07
圧勝	0.07
勝ち	0.16

4. 実　　験

提案手法の有用性を確認するために，評価実験を行った．

4.1 実 験 手 法

本実験では，Web ニュース 50 記事を用いて，各記事のタイトルに対して提案手法により例え表現を生成した．実験に用いた Web ニュース記事には，政治，社会，スポーツ，芸能等の各カテゴリを広く取り入れ，なおかつ記事の重要性や価値が評価の指標となるため，評価者全員が記事の内容をすでに把握している過去 4 年以内の主要なニュースを対象とした．例え表現生成時のユーザの興味として「野球」と「ゲーム」の 2 つを入力し，野球に関する例え表現 50 文,「ゲーム」に関する例え表現 50 文の計 100 文の例え表現を生成した．次に，生成した例え表現に関して,「野球」に興味を持つ 20 代の男性評価者 3 名と「ゲーム」に興味を持つ 20 代の男性評価者 3 名の計 6 名がニュースの例えとして妥当かどうかの評価を行った.「野球」に興味を持つ評価者 3 名は「野球」の例え表現 50 文,「ゲーム」に興味を持つ評価者 3 名は「ゲーム」の例え表現 50 文をそれぞれ評価した．評価値は，例えとして妥当ではない時は 0，例えとしてある程度妥当であれば 1，例えとして妥当であれば 2 とした．

表 4　例え表現の評価結果

興味	生成した例え表現数	妥当と評価した数	適合率
野球	50	28	0.56
ゲーム	50	27	0.54
野球＋ゲーム	100	55	0.55

4.2 実験結果と考察

本実験では,「野球」および「ゲーム」のそれぞれ評価者 3 人の評価値の合計が 3 以上となった例え表現を例え表現として妥当と判別した．表 4 に「野球」および「ゲーム」に関して生成した例え表現の内，妥当と評価した数とその適合率を示す．提案手法により生成した例え表現 100 文に対して 55 文が妥当という評価となった．適合率は 0.55 で 0.5 を超えており，提案手法による例え表現生成が有用であることが確認できる．また,「野球」と「ゲーム」それぞれの適合率は 0.56，0.54 となり，両方共に 0.5 を超えた.「野球」,「ゲーム」共に評価に大きな差が見られなかったことからも，提案手法では異なる興味を指定した場合においても十分に例え表現を生成することが可能であると考えられる．

野球の例え表現

ユーザの興味を「野球」で生成した例え表現として評価値が高かった例え表現の一部を表 5，評価値が低かった例え表現の一部を表 6 にそれぞれ示す．3 人の評価者全てが妥当と評価したのは,「本田圭佑が AC ミランへ移籍」に対しての「田中将大がヤンキースへ移籍」のみであった．この例え表現に関しては田中将大は実際にニューヨーク・ヤンキースに移籍しており，今回生成した例え表現の中では唯一事実に基づいたものとなった．逆に「福山雅治が吹石一恵と結婚」に対して「イチローが山本昌と結婚」や「堀北真希が山本耕史と結婚」に対する「金田正一が田口壮と結婚」のように実際に起こりえない事象に関しては評価が低いものとなっている．本論文における例え表現では，先にも述べたように生成する表現の自由性を求める目的で例え表現の内容が事実であるかは問わないとしてきたが，事実に基づく例え表現の方がよりユーザに受け入れ易く，理解に繋がるのではないかと考えられる．生成する例え表現に関して事実性を考慮し Web 上からも積極的に広く情報を抽出する必要があると考えられる．また，評価の高かった例え表現では，元のニュースが,「本田圭佑が AC ミランへ移籍」,「ラグビー日本代表が南アフリカに歴史的勝利」,「吉田沙保里が国民栄誉賞

表 5　高評価を得た「野球」の例え表現

ニュースタイトル	例え表現（野球）	評価値合計
本田圭佑が AC ミランへ移籍	田中将大がニューヨーク・ヤンキースへ移籍	6
又吉直樹が芥川賞を受賞	板東英二が最多安打を獲得	5
ラグビー日本代表が南アフリカに歴史的勝利	阪神タイガースがロサンゼルス・ドジャースに大勝	5
ジョージ・ルーカス監督が「フォースの覚醒」を批判	ホセ・カンセコがミスター・ベースボールを批判	5
吉田沙保里が国民栄誉賞を受賞	王貞治が IBM プレイヤー・オブ・ザ・イヤー賞を受賞	5

表 6　低評価を得た「野球」の例え表現

ニュースタイトル	例え表現（野球）	評価値合計
福山雅治が吹石一恵と結婚	イチローが山本昌と結婚	1
堀北真希が山本耕史と結婚	金田正一が田口壮と結婚	0
ニューホライズンズが冥王星を撮影	伊藤博史が佐藤二朗をロケ	0
アメリカがキューバとの国交回復	セントラル・リーグが日本プロフェッショナル野球組織との停戦	0
イスラエルがハマスと停戦	日本野球規則委員会がプロ野球 28 会で反乱	0

表 7　高評価を得た「ゲーム」の例え表現

ニュースタイトル	例え表現（ゲーム）	評価値合計
DeNA が任天堂と資本提携	ソニー・コンピュータエンタテインメント が任天堂と提携	6
又吉直樹が芥川賞を受賞	NHK エンタープライズが日本ボードゲーム大賞を受賞	5
大村智がノーベル賞を受賞	セガ・インタラクティブ が日本ゲーム大賞を獲得	5
松山英樹が PGA ツアーで初優勝	PHP 研究所が日本ゲーム大賞で受賞	5
天空の城ラピュタが金曜ロードショーで放送	マルサの女が東京ゲームショーで配信	5

表 8　低評価を得た「ゲーム」の例え表現

ニュースタイトル	例え表現（ゲーム）	評価値合計
福山雅治が吹石一恵と結婚	任天堂が SNK プレイモアと和解	1
avex が JASRAC を脱退	エンターブレインが任天堂を退社	1
イギリスのウィリアム王子に長女誕生	ありす in Cyberland のスクウェア・エニックスに母	0
ＡＳＫＡが覚せい剤で逮捕	メディアファクトリーが覚醒剤で逮捕	0
高倉健が悪性リンパ腫により死去	タカラトミー が悪性リンパ腫により病死	0

を受賞」等のようにそれぞれ「サッカー」,「ラグビー」,「レスリング」に関するニュースであり広義には「スポーツ」に関するニュースとなっている．指定したユーザの興味の「野球」も「スポーツ」に属するものである．今回例え表現に用いたニュース記事 50 記事の内 16 記事が「スポーツ」に関するニュースであったが，その内妥当と評価した例え表現は 12 文あり,「スポーツ」記事だけで評価した場合，適合率は 0.75 と 50 記事に対しての評価よりさらに高い値となる．このことから，例え表現生成では，指定するユーザの興味が例えるニュース記事と広義にカテゴリが一致していることが望ましいと考えられる．

一方で,「S（主語）」,「O（目的語）」,「V（述語）」それぞれの変化に着目すると，高評価を得た例え表現では「本田圭佑→田中将大」,「ラグビー日本代表→阪神タイガース」,「国民栄誉賞→ IBM プレイヤー・オブ・ザ・イヤー賞」のように選手には選手，チームにはチーム，賞には賞で S および O を例えている．これに対して低評価を得た例え表現では,「冥王星→佐藤二朗」,「アメリカ→セントラル・リーグ」のように惑星に対して人物，国に対してリーグ等カテゴリの一致性が弱い S と O を抽出している．このことからも例え表現の生成では，例えるもののカテゴリが一致しているもので例えることが相応しいと考えられるが，例えるものの語やユーザの興味によっては，必ずしも一致するカテゴリを抽出することはできない．そこで，カテゴリの意味を保持しながら広義に解釈し置き換える必要がある．今後は，カテゴリ間の関係構造をグラフ化し，ノード間の距離関係から適切なカテゴリの抽出を試みたい．

V（述語）に関しては，高評価を得た例え表現と評価を得た例え表現共にもとのニュースタイトルから変化していないものも多い．男性野球選手同士の「結婚」や「停戦→反乱」のように低評価を得た例え表現では文として意味が破綻していたり，もとのタイトルの意味から離れている場合が多い．例え表現では，文全体の意味を考慮することも重要であるため，今後は S に係る V の抽出を優先して，抽出した S と V の関係から O を抽出する流れでの例え表現生成手法を確立し，今回の手法との結果の比較を行いたいと考えている．

ゲームの例え表現

ユーザの興味を「ゲーム」で生成した例え表現として評価値が高かった例え表現の一部を表 7，評価値が低かった例え表現の一部を表 8 にそれぞれ示す.「ゲーム」の例え表現で高い評価を得たものには，ニュースタイトルの各賞の受賞に対してゲームの賞の受賞で例えたものが多かった．これは，先に述べたようにカテゴリが一致していることに加え，賞の例えは価値の度合いが直感的に把握しやすいためと考えられる．逆に，人物の

例えでは,「高倉健→タカラトミー 」,「福山雅治→任天堂」のように「ゲーム」では企業等で例えていることが多く，人物で例えていないため結果的に文の意味が破綻して評価が低くなっている．また,「野球」に関しても同様だが「国」に関する例えは低評価を得ることが多かった．ニュースの題材には，人物や国に関する事柄が頻繁に出現することからもニュースの題材に成りやすい特定のカテゴリに関してはカテゴリ間の繋がりを機械学習することが例え表現生成の精度向上に繋がると考えられる.「ゲーム」の例え表現では,「DeNA が任天堂と資本提携」を「ソニー・コンピュータエンタテインメント が任天堂と提携」と例えたように，会社や作品に関して「野球」で妥当と評価しなかったニュース記事タイトルに関しても妥当と評価している．一方で,「スポーツ」関連の記事に関しては,「野球」程の評価は得られなかった．ユーザの興味によってある程度例え表現の生成精度が異なることがわかる．指定するユーザの興味が例えるニュース記事と広義にカテゴリが一致していることが望ましいのがわかったため，例えるニュース記事に対してシステム側でユーザの興味を拡張，または推薦する前処理が必要と考えられる．

5. まとめと今後の課題

本論文では，Web ニュースの理解支援を目的として，ユーザが閲覧しているニュース記事のタイトルをユーザの興味のある内容で例える例え表現の生成自動生成手法を提案した．例え表現は，ユーザの興味と例えるものの価値 2 つを考慮して行った．ユーザの興味では，Wikipedia のカテゴリ構造を利用してユーザの入力する興味に柔軟に対応して例え表現に用いるための主語や目的語を抽出する手法を提案した．ものの価値については，価値の指標として 2 つの仮説を立て PageRank アルゴリズムと Wikipedia の閲覧回数を用いて単語の価値を推定する手法を提案した．また評価実験を行い，生成した例え表現が適合率 0.55 で妥当と評価され，提案手法の有用性を確認した．

今後の課題として，カテゴリ構造のより適切な抽出手法および文脈を考慮した例え表現の生成手法の考案が課題である.「S, O, V の形式以外のニュース記事や Wikipedia に情報がない語の価値の推定手法の確立も必要と考えられる．

謝辞

本論文の一部は 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．

文　献

[1] 新谷研, 角田達彦, 大石巧, 長尾眞, “単語の共起頻度と出現位置による新聞の関連記事の検索手法”, 情報処理学会論文誌, Vol. 38, No. 4, pp. 855–862, apr 1997.

[2] 奥村学, 難波英嗣, “テキスト自動要約に関する研究動向”, 自然言語処理, Vol. 6, No. 6, pp. 1–26, 1999.

[3] 真下遼, 灘本明代, “Web ニュース内容理解支援のための例え表現自動生成手法の提案 (データ工学)”, 電子情報通信学会技術研究報告 = IEICE technical report : 信学技報, Vol. 115, No. 177, pp. 91–95, aug 2015.

[4] 平田紀史, 白松俊, 大囿忠親, 新谷虎松, “ユーザの観点に基づくイベント系列化を用いた web ニュース記事閲覧支援システムの実装”, 人工知能学会論文誌, Vol. 26, No. 1, pp. 228–236, 2011.

[5] 張建偉, 河合由起子, 熊本忠彦, 白石優旗, 田中克己, “多様な印象に基づくニュースサイト報道傾向分析システム”, 知能と情報, Vol. 25, No. 1, pp. 568–582, 2013.

[6] 石井裕志, 馬強, 吉川正俊, “Svo 構造を用いた因果関係ネットワーク構築手法について”, 研究報告データベースシステム（DBS）, Vol. 2009, No. 10, pp. 1–8, nov 2009.

[7] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song, “Newscube: Delivering multiple aspects of news to mitigate media bias”, In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’09, pp. 443–452, New York, NY, USA, 2009. ACM.

[8] 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.

[9] Satoshi Oyama and Katsumi Tanaka, “Query modification by discovering topics from web page structures”, Vol. 3007, pp. 553–564, 2004.

[10] 西原陽子, 砂山渡, 谷内田正彦, “Web ページの難易度と学習順序に基づく情報理解支援システム (コンテンツ技術,web 情報システム)”, 電子情報通信学会論文誌. D, 情報・システム, Vol. 89, No. 9, pp. 1963–1975, sep 2006.

[11] 佃洸摂, 大島裕明, 田中克己, “上位下位概念辞書を用いた同位語・上位語のランキング手法の提案”, Web とデータベースに関するフォーラム (WebDB Forum 2013, 2013.

[12] 乾健太郎, 藤田篤, “言い換え技術に関する研究動向”, 自然言語処理, Vol. 11, No. 5, pp. 151–198, 2004.

[13] 木虎直樹, 久保征人, “Web アクセス履歴に基づくユーザの価値観の類推”, 人工知能学会全国大会論文集, Vol. 27, pp. 1–3, 2013.

[14] 奥健太, 中島伸介, 宮崎純, 植村俊亮, 加藤博一, “情報推薦におけるユーザの価値判断基準モデルに基づくコンテキスト依存型ランキング方式”, 情報処理学会論文誌データベース（TOD）, Vol. 2, No. 1, pp. 57–80, mar 2009.

[15] Sergey Brin and Lawrence Page, “The anatomy of a large-scale hypertextual web search engine”, *Comput. Netw. ISDN Syst.*, Vol. 30, No. 1-7, pp. 107–117, April 1998.

[16] Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio, “Wikipedia mining for an association web thesaurus construction”, In *Web Information Systems Engineering – WISE 2007*, Vol. 4831 of *WISE 2007*, pp. 322–334, 2007.