Vol.2018-EC-50 No.31
2018/12/22

ゲーム用語に着目したWord2vecを用いたゲームジャンルの推定精度評価

大山 浩暉[†1,a)]　竹川 佳成[†1,b)]　平田 圭二[†1,c)]

概要：ユーザがプレイしたいゲームを検索エンジンを用いて探索することは時間の浪費であり，正確なゲーム情報を取得するのに不十分な面がある．本研究では，ジャンルに着目し，Word2vec を用いた正確なゲーム情報をピックアップすることを目的とする．実験では，ゲーム用語に焦点を置き，重みづけを行った学習データを用いて Word2vec に学習させ，その出力と入力の合致度を正解率として比較，評価を行った．学習データは 4 つあり，「レビュー文から形態素解析し，重みづけを行ったデータ」，「レビュー文から名詞，動詞，形容詞のみ抽出し，重みづけを行ったデータ」，「レビュー文から形態素解析し，重みづけを行ったデータに wiki のテキストデータを加えたデータ」，「レビュー文から名詞，動詞，形容詞のみ抽出し，重みづけを行ったデータに wiki のテキストデータを加えたデータ」を Word2vec に学習させた．結果はそれぞれ，0.64，0.7，0.69，0.73 となった．

1. はじめに

ユーザが自分がプレイしたいゲームを探索するとき，検索エンジンを利用することが一般的である．そして，検索エンジンで検索を行う際に用いられるキーワードとして，ゲームのジャンルが挙げられる．しかし，現在の検索エンジンでは，ゲームのジャンルを用いて検索を行っても，そのジャンルに合致したゲーム情報を正確に取得することができない．

表 1 では，「アドベンチャーゲーム WiiU」で検索した結果として出力されたサイト 10 件分である．この表にあるゲームサイトのほとんどは，アドベンチャーゲームに相当すると思われるゲームを 20 件以上羅列したサイトとなっている．この中から自分に合ったゲームを探し出すのは時間の浪費である．加えて，アドベンチャーゲームというジャンルに合致していないゲームも結果として表示されている．

上記の問題を解決するためには，ゲームのジャンルに対してジャンルと合致したゲームをピックアップする工夫が必要である．本研究の目的は，ゲームジャンルによるゲーム推定の精度を向上することである．ゲームのジャンルから正確なゲームを機械学習で出力させるために，学習データを作成する．その作成した学習データを評価し，考察を行う．

表 1 ”アドベンチャーゲーム WiiU”の検索エンジンの結果

検索結果
トマトアドベンチャー — Wii U — 任天堂...
Wii おすすめアドベンチャーゲームソフト - ...
Amazon.co.jp: アドベンチャー - ゲームソフト...
モモンガ ピンボール アドベンチャー — Wii U...
WiiU のジャンル「アドベンチャー」のゲーム一覧...
アドベンチャー 01 - Wii U コレクション
GAME ダッシュ - アドベンチャーゲーム（WiiU...
wiiU アドベンチャーゲーム一覧
Wii U おすすめアドベンチャーゲームソフト -
Wii U アドベンチャーゲーム ダウンロード 通販

2. 関連研究

2.1 Word2vec

Word2vec とは，入力されたテキストデータから単語をベクトル化して表現する定量化手法である [1]．各単語を任意の低次元ベクトル空間上（例えば 200 次元空間など）で表現し，それらを用いて単語の類似度計算や意味演算を可能としている．Word2vec は Mikolov らによって提案されたニューラルネットワークを用いた手法で，単語の特徴をとらえた単語ベクトルを学習する [2]．複数の単語列から構成されるコーパスを入力として受け取っており，入力層における語彙数の次元を低次元ベクトル空間における単語ベクトルとして表現することを目的としている．Word2vec は同じような文脈情報から単語ベクトルの学習を行い，そ

†1 現在，公立はこだて未来大学大学院
a) g2117010@fun.ac.jp
b) yoshi@fun.ac.jp
c) hirata@fun.ac.jp

ⓒ 2018 Information Processing Society of Japan

こで得られた単語ベクトルは，単語の特徴を表した分散表現と呼ばれる．Word2vec では，ある単語の近くに存在する単語は似た意味を持つという考えの元，近くに存在する単語は関係性が強いと判断され，類似度が高くなる．

本研究は学習データとしてゲームのレビュー文を用いるが，レビュー文にも適応できると考える．例えば，「感動できるエンディングに仕上がってます」や，「謎解きパズル要素やミニゲームも楽しい」といったことがレビュー文に記述されている．これらは感動とエンディングは同等の意味を持ち，謎解きパズルと楽しい，ミニゲームと楽しいも同等の意味を持つと考えられる．これらはお互いに近くに存在するため，Word2vec で学習させる際，お互いの類似度を高く学習できるため，本研究では Word2vec が最も適した技術であると考える．そのため，本研究では，Word2vec を用いて作成した学習データを学習させ，精度を評価，考察を行う．

2.2 ゲーム推薦

ゲーム推薦に関連する研究として，中谷らのユーザの経験的価値を元にゲーム推薦を行う研究がある [3]．中谷らは，個人向け電子商取引のウェブサイトでは，推薦システムが導入されているが，ゲームの重要な要素である経験的価値を考慮していないことを問題点として指摘している．この経験的価値というのは，ゲームをプレイした経験がある人にしか説明することができない感想を意味する．ゲームに経験的価値を基に求められた経験値が付与されているものとして，経験値がゲームからユーザに移動すると考えた．中谷らの研究では，経験的価値は感覚的知覚，創造的思考，身体性，社会性，衝動的感情の 5 つに分類されている．協調フィルタリングに基づく推薦システムでは他社の購入履歴などの情報からユーザの嗜好を予測し推薦するが，ゲーム同士の類似性を具体的に得ることは難しい．一方で推薦システムの研究では，内容ベースフィルタリングに基づく推薦システムが多く提案されている．本やニュース記事といった対象は，テキスト情報と画像から構成されることが多いため，有効な特徴量を抽出することは比較的容易であるが，ゲームのようなマルチメディアコンテンツでは特徴量を抽出することは困難である．そこで，中谷らは，レビューサイトにおける経験的価値を含んだ文章の多用さに着目し，文章の抽出を辞書を用いてテキストマイニングでゲームの分類を行った．そして，ゲームの経験を定量化し経験値へと変換させ，各ユーザの経験に適するゲームを推薦するシステムを提案した．本研究では，ゲームを経験的価値から推定するのではなく，ジャンルから推定する．そのため，ほとんどのジャンルを網羅している最も少ないジャンル数を調査し，それぞれのジャンルを定義した．そのジャンルからゲームを推定し，評価を行う．

2.3 専門用語抽出

専門用語を抽出する方法として中川らの出現頻度と連接頻度に基づいた専門用語を抽出する研究がある [4]．中川らは名詞（単名詞と複合名詞）を対象として専門用語抽出について検討した．単名詞のバイグラムから得られる短名詞の統計量を利用し，ある単名詞が複合名詞を形成するために連接する名詞の頻度を用いた．この頻度を利用した数種類の複合名詞スコア付け法を提案した．中川らの研究では，単名詞と複合名詞に着目し，複合名詞は連続した名詞を抽出している．しかし，本研究ではゲームのレビュー文から用語を抽出するが，口語が多く含まれているため連続した名詞の抽出ではうまくいかないと考えた．例えば，「あまり」と「関係」や「♪」と「他」が連続した名詞となっていた．ゲームレビュー文は一般の人が記述してあり，記述方法にも制限がないため，口語や記号等が多く含まれているため，連続した名詞の抽出ではうまくいかないと考えられる．

他の専門用語抽出研究として，木浪らの看護学分野の専門用語抽出方法の研究がある [5]．木浪らは，産学連携の専門家が研究のシーズを専門用語によって簡単に検索することができるシステムの構築を狙いとし，その第一段階として専門用語抽出の研究を行っていた．木浪らの研究では，これまで研究されていない看護学分野を対象分野とし，専門用語になりうる品詞組み合わせの拡張と一般的な語を除去することで専門用語抽出の性能改善を図った．木浪らは専門用語を抽出する際，専門用語として正しいかを測る正解データを看護学専門家のもと作成した．しかし，ゲームという分野では，専門用語抽出の正解データ作成に専門家を用いることはできない．なぜなら，ゲームは一般的な用語が多く含まれており，ゲームにしか使用されない専門用語の定義がなく，専門用語が記述されている辞書といったものも存在せず，ゲームの専門家も存在しないためである（ゲーム開発者はゲームの専門家というより，プログラミングやデザイナーの専門家であると考えられる）．以上から，ゲームにしか存在しない専門用語を抽出することは困難であるため，本研究ではゲームの専門用語ではなく，ゲームでよく扱われる単語に着目した．

3. 学習データの作成

3.1 特徴量の選定

学習データの元となる文書はゲームレビューサイト mk2 とファミ通.com を使用した．本サイトを使用した理由は 2 つある．1 つは他のサイトと比較した際，1 レビュー文あたりの情報量が多く，質も高いためである．1 レビュー当たりの文字数が 100 文字を超える記述がされており，ゲーム以外の不必要な情報が記述されていないレビュー文が多く見られた．2 つ目はゲームをプレイするプレイヤが感じたゲームの特徴が記述された文書を用いたいからである．

ゲームのホームページにはゲーム開発者が記述したゲームの概要がある．しかし，そのゲームに対するゲーム開発者にとって良い点とプレイヤにとって良い点は異なる場合がある．そのため，本研究では，プレイヤを想定し，ゲーム情報を取得する．このゲームレビューサイトから抽出したゲーム名の数は 7299 で，レビュー数は 30429 である．

3.2 ゲーム用語の抽出

ゲームレビュー文を分析し，レビュー文の特徴を調査した．オンラインとオフラインによる感想傾向の違いや，レビュー文記述者と同等の立場に伝える形式が多い傾向などが見つかった．その中で特に特徴が出ていると思われる要素を仮説として立てた．それは，詳細に記述されているレビュー文はゲームでよく扱われるような単語や表現を用いているという仮説である．

以上の特徴から，レビュー文で良く扱われる単語をゲーム用語とし，ゲーム用語に重みづけを行ったレビュー文を用いることで，ジャンルによるゲーム推定の精度が向上すると考えられる．

N-gram を用いて元のレビュー文から N 文字を抽出した．N-gram を用いた理由は，形態素解析では分割されてしまう単語を取得するためである．例えば，レビュー文でよく扱われる単語の「難易度」や「全体的」といった単語は，形態素解析によって「難易」と「度」,「全体」と「的」に分割されてしまう．これらの単語は分割されず，1 つの単語として処理したいため，今回は N-gram を用いた．N-gram を用いて抽出した単語はゲーム用語として扱い，形態素解析時にユーザ辞書としても用いた．

抽出方法は図 1 に示す．今回は，N は文字数と設定したため，N=3 は 3 文字を表している．例えば，「面白いゲームです」といった文に N=3 として N-gram を適応した場合，「面白い」,「白いゲ」,「いゲー」,「ゲーム」,「ームで」,「ムです」という単語が取得される．加えて，N-gram で得た単語を絞り込むため，条件に合致した単語のみを抽出した単語の集合に加えた．抽出する条件は 5 つある．単語が 3 文字から 7 文字である，記号が含まれていない，数字のみの単語になっていない，意味不明な未知語ではない，出現回数が 200 以上である．2 文字以下はゲーム用語として設定するには単語の量が過多であった．8 文字以上は意味を持たない単語が多く，本来の単語から欠損しているものも多かったためである．記号や数字のみの単語には意味を持たないと考え，今回は排除した．N-gram を用いると意味を持たない未知語が多く存在した．例えば「ょうがない」や「オリジナリ」のような単語の一部が欠損したものである．以上で得たゲーム用語の数は 1582 語である．N=3 は 1053 語，N=4 は 325 語，N=5 は 143 語，N=6 は 44 語，N=7 は 17 語である．

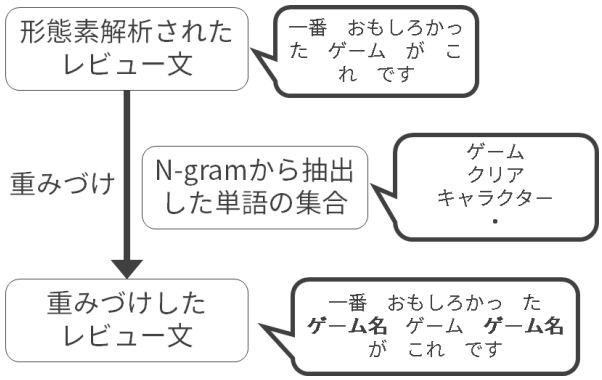

図 1 ゲーム用語の抽出方法

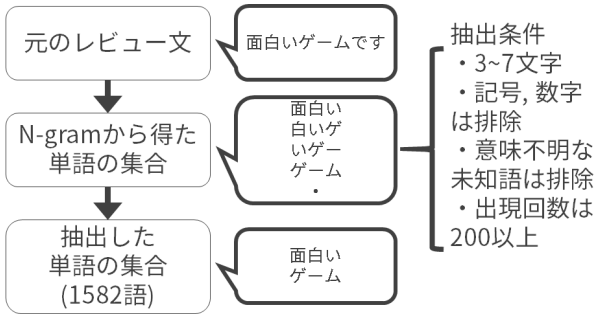

図 2 重みづけ学習データの作成方法

3.3 レビュー文の重みづけ

今回行った重みづけとは，元のレビュー文から MeCab を用いた形態素解析を行い，N-gram から抽出した単語が元のレビュー文中に存在するか探索し，もし存在した場合，その単語の周りに，レビュー文に対応するゲームの名前を左右に 1 つずつ挿入することを示す [6]．例として，N-gram を用いてゲーム用語に重みをつける方法を図 2 に示す．形態素解析されたレビュー文に N-gram から抽出した単語が存在するか探索した．例では，N-gram から抽出した単語の集合に「ゲーム」が含まれており，形態素解析されたレビュー文にも「ゲーム」が含まれている．そのため，形態素解析されたレビュー文にある「ゲーム」の左右に，レビュー文に対応したゲーム名を付与した．この作業で作成したデータを重みづけされたレビュー文データとした．

Word2vec はある単語に近い単語は似た意味を持つと考え，類似度が高くなる．そのため，抽出したゲーム用語にゲームの名前を近くに付与することで，レビュー文とゲーム名の関係性を強くする重みづけが可能となる．

以上でできたレビュー文を学習データとして Word2vec の学習に使用する．

3.4 wiki のテキストデータの追加

データ量を増加させることでジャンルによるゲーム推定の精度が向上すると考えられる．しかし，元のレビュー文のみでは 77.6MB しかないためデータ量が不足していた．他レビューサイトのレビュー文抽出も試みたが，ページ

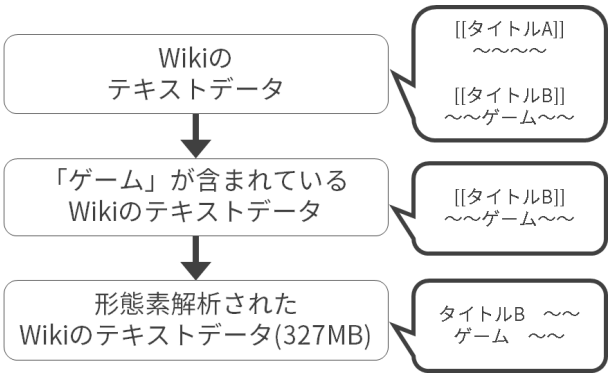

図 3 wiki のテキストデータの抽出方法

URL の不規則性などで，ゲームレビュー文のみを取得することが困難であった．そのため，他のテキストデータを用意する必要がある．そこで，wiki のテキストデータを用い，「ゲーム」という単語が含まれるページのみを抽出した．抽出方法を図 3 に示す．ゲームが含まれるページのみの Wiki テキストデータは 327MB であり，このデータも学習データに加える．

4. 実験の内容

本実験では，Word2vec を用いてゲームのジャンルを入力とし，出力されるゲームの推定精度を比較，評価する．

ジャンルの数は，全てのゲームを網羅する最低数であろう 8 つに定めた．8 つのジャンルとそれらの定義を表 2 に示す．ジャンルは著者を含めた 2 名が Wikipedia の当該ページを参考に定義した

対象データは「元レビュー文の形態素解析のみ＋ N-gram で得た単語の重みづけ」，「元レビュー文の形態素解析から名詞，動詞，形容詞のみ＋ N-gram で得た単語の重みづけ」，「それぞれに形態素解析のみ行った wiki データ」の 4 つであり，それぞれを学習させた結果を比較する．ゲームの特徴を表すレビュー文は形態素解析のみで十分か，名詞，動詞，形容詞により着目した方が良いのかを確かめ，同時に Wiki のテキストデータを加えることによる精度向上が見られるのかを本実験で確認する．

評価の方法は図 4 に示す．学習データを Word2vec に学習させ，ゲームと入力との類似度が高い上位 10 件のゲームを出力させた．その出力が合致しているかを各ゲームのホームページ等から手動で判断した．判断基準は，ホームページにあるゲームの概要，ジャンル名に入力と同じ単語が含まれているか否かで判断した．

5. 結果

学習データ「元レビュー文の形態素解析のみ＋ N-gram で得た単語の重みづけ」と，「元レビュー文の形態素解析から名詞，動詞，形容詞のみ＋ N-gram で得た単語の重みづけ」の 2 つの結果を表 3 に示す．

表 2 ゲームのジャンルと定義

ID	ジャンル名	定義
(1)	シューティング	敵自身や敵の攻撃を避けながら射撃武器で敵や的を狙い撃つ
(2)	アクション	ゲーム内でプレイヤがキャラクタをコントローラで操作し，リアルタイムに起こる様々な課題や障害を攻略する
(3)	アドベンチャー	物語上に配置された謎を解く
(4)	RPG	課題や障害を解決した報酬としてプレイヤの操作キャラクタを強化し，物語のエンディングを目指す
(5)	パズル	出題された問題を試行錯誤によって解く
(6)	レース	乗り物を操作し，他のプレイヤやコンピュータと競争する
(7)	シミュレーション	プレイヤがある役割となって戦略的に目標を達成するあるいは，自分の駒を動かして相手の駒を倒す
(8)	音楽	リズムや音楽に合わせてボタンを操作する

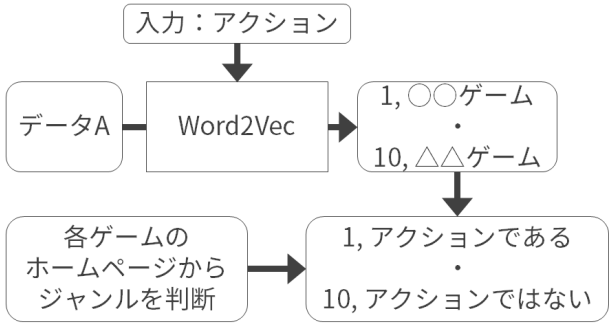

図 4 実験の評価方法

正解率とは，入力であるジャンル名と出力であるゲーム名が合致している割合を示す．例えば，0.5 は 10 件中 5 件のゲームは入力と合致していたことを示す．共通しているのは，シミュレーションのジャンルの正解率が低い傾向にあった．他は RPG や音楽も低い傾向にあった．

上記 2 つの学習データそれぞれに形態素解析のみ行った wiki データの 2 つの結果を表 4 に示す．共通しているのは，シミュレーションの正解率が低い傾向にあった．他は RPG も低い傾向にあった．

形態素解析のみを行った場合と，形態素解析から名詞，動詞，形容詞のみを抽出した場合，後者の方が正解率が上がる傾向にあった．加えて，形態素解析された wiki のデータを付与することで，より正解率が上がる傾向にあった．つまり，wiki を加え，名詞，動詞，形容詞のみのレビュー文を用いることで正解率が高くなることが分かった．しかし，シミュレーションはどのパターンにおいても変わらず低いままとなった，RPG に関しても低いものがあった．

6. 考察

今回の実験では n-gram を用いてゲームでよく扱われるような単語を抽出し，それらの単語に重みを付けたレビュー

表 3　wiki 無しデータの正解率

ジャンル名	形態素解析 +重みづけ済	名詞，動詞，形容詞 +重みづけ
シューティング	0.7	0.7
アクション	1.0	1.0
アドベンチャー	0.9	0.7
RPG	0.4	1.0
パズル	0.7	0.7
レース	1.0	0.9
シミュレーション	0.3	0.4
音楽	0.5	0.2
平均正解率	0.64	0.7

表 4　wiki 有りデータの正解率

ジャンル名	形態素解析 +重みづけ済+wiki	名詞，動詞，形容詞 +重みづけ+wiki
シューティング	0.5	0.7
アクション	1.0	1.0
アドベンチャー	0.8	0.8
RPG	0.5	0.3
パズル	0.9	0.9
レース	1.0	1.0
シミュレーション	0.2	0.4
音楽	0.6	0.7
平均正解率	0.69	0.73

文を作成し，そのデータを Word2vec の学習データとして扱った．しかし，N-gram の他に TF-IDF を用いたゲーム用語の抽出が考えられる．TF-IDF により，N-gram で得たゲーム用語をより質の良い単語として抽出できることが期待できる．そのため，N-gram を用いたゲーム用語抽出方法と TF-IDF を用いたゲーム用語抽出方法で比較実験を行う必要がある．比較実験は今回行われた実と同様にし，正解率を比較する実験を行う予定である．

今回実施できなかった内容として，否定文の考慮が挙げられる．今回用いた学習データでは，否定文に対しての処理は行っていない．レビュー文にはゲームの良かった点だけでなく，悪かった点や改善点を述べていることが多い．そのため，否定文を考慮することで，ゲーム名とレビュー文の関係性がより強くなった学習データを作成することが可能となる．否定文を考慮する方法として，Cabocha の係り受け解析と小林らの日本語評価極性辞書（用言編）や東山らの日本語評価極性辞書（名詞編）を用いることで可能だと考える [8], [9]．加えて，「ない」や「ありません」といった否定を表す語が出現した場合，極性を反転させることや，係り受け関係にある単語を排除するといった処理をすることで，否定語を考慮した処理が可能と考える．

7.　今後の展望

今回の実験では n-gram を用いてゲームでよく扱われるような単語を抽出し，それらの単語に重みを付けたレビュー文を作成し，そのデータを Word2vec の学習データとして扱った．しかし，N-gram の他に TF-IDF を用いたゲーム用語の抽出が考えられる．TF-IDF により，N-gram で得たゲーム用語をより質の良い単語として抽出できることが期待できる．そのため，N-gram を用いたゲーム用語抽出方法と TF-IDF を用いたゲーム用語抽出方法で比較実験を行う必要がある．比較実験は今回行われた実と同様にし，正解率を比較する実験を行う予定である．

今回実施できなかった内容として，否定文の考慮が挙げられる．今回用いた学習データでは，否定文に対しての処理は行っていない．レビュー文にはゲームの良かった点だけでなく，悪かった点や改善点を述べていることが多い．そのため，否定文を考慮することで，ゲーム名とレビュー文の関係性がより強くなった学習データを作成することが可能となる．否定文を考慮する方法として，Cabocha の係り受け解析と小林らの日本語評価極性辞書（用言編）や東山らの日本語評価極性辞書（名詞編）を用いることで可能だと考える [7], [8], [9]．加えて，「ない」や「ありません」といった否定を表す語が出現した場合，極性を反転させることや，係り受け関係にある単語を排除するといった処理をすることで，否定語を考慮した処理が可能と考える．

表 5　レビュー文と wiki テキストデータのジャンル出現回数

ジャンル名	レビュー文中の出現回数	wiki 中の出現回数	出現回数の合計
シューティング	2134	5263	7397
アクション	12525	22065	34590
アドベンチャー	7908	12852	20760
RPG	10007	11370	21377
パズル	5298	7613	12911
レース	4916	11722	16638
シミュレーション	790	4226	5016
音楽	11371	29192	40563

参考文献

[1] 西尾泰和: Word2vec による自然言語処理, オライリー・ジャパン （2014）.
[2] T.Mikolov, K. Chen, G. Corrado, and J. dean: Efficient Estimation of Word Representations Vector Space. In ICLRWorkshop （2013）.
[3] 中谷知博, 星野雄一: 経験的価値の分類に基づくゲーム推薦システム, 一般社会法人 情報処理学会 情報処理学会研究報告エンタテインメントコンピューティング, 2008-EC-011, Vol.2008, No.129, pp.49-56 （2008）.
[4] 中川裕志, 湯本紘彰, 森辰則: 出現頻度と連接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）.
[5] 木浪孝治, 池田哲夫, 村田嘉利, 高山毅, 武田利明: 看護学分野の専門用語抽出方法の研究, 自然言語処理, Vol.15, No.3, pp.3-20 （2008）.
[6] Mecab: Yet Another Part-of-Speech and Morphological Analyzer(online), 入手先 ¡http://taku910.github.io/mecab/¿ (2018.11.29).
[7] Cabocha/南 瓜: Yet Another Japanese Dependency Structure Analyzer(online), 入手先 ¡http://taku910.github.io/cabocha/¿ (2018.11.29).

[8] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一: 意見抽出のための評価表現の収集, 自然言語処理, Vol. 12, No. 2, pp. 203-222 （2005）.
[9] 東山昌彦, 乾健太郎, 松本裕治: 述語の選択選好性に着目した名詞評価極性の獲得, 言語処理学会第 14 回年次大会論文集, pp. 584-587（2008）.

ⓒ 2018 Information Processing Society of Japan