ブログ記事からのオノマトペ用例文の自動抽出手法[†]

内田 ゆず^{*1}・荒木 健治^{*2}・米山 淳^{*1}

日本語の語彙には、オノマトペが豊富に存在している。これらの語は生き生きとした表現力をもち、日本語でのコミュニケーションには欠かせないものとなっている。日本語を母語とする人は、ごく自然にオノマトペの用法を身につけ、「感覚的に」使用する。そのため、オノマトペをほかの言葉に置き換えたり、その意味を明確に説明することは困難である。日本語学習者がオノマトペの意味や用法を習得することは難しいとされている。オノマトペを学ぶ方法のひとつとして、オノマトペを用いた様々な表現に触れることが挙げられる。そこで、本論文では、ブログ記事を対象としたオノマトペ用例文の自動抽出手法を提案する。提案手法は、オノマトペの後続要素と係り先を分析した結果に基づくものである。評価実験の結果、41,315件のブログ記事から15,437文の用例文を抽出し、適合率96.2% (481/500) が得られた。

キーワード：オノマトペ、後続要素、構文解析、コーパス

1. はじめに

日本語の語彙には、動作や状態、事物の姿形を感覚的に表す擬態語、事物の音や声を表す擬音語が豊富に存在している。（以降、擬態語・擬音語をまとめて“オノマトペ”と呼ぶ。）これらの語は生き生きとした表現力をもち、日本語でのコミュニケーションには欠かせないものとなっている。日本語を母語とする人は、ごく自然にオノマトペの用法を身につけ、「感覚的に」使用する。そのため、オノマトペをほかの言葉に置き換えたり、その意味を明確に説明することは困難である。

オノマトペは日常生活の中で多用されるが、日本語教育の現場ではその使用頻度や重要性に見合うだけの十分な学習や指導がなされているとはいえない[1]。日本語学習者がオノマトペの意味や用法を習得することは難しく、有効な指導法・学習法の確立が求められている。

オノマトペを学ぶ方法の一つとして、オノマトペを用いた様々な表現に触れることが挙げられる。これを実践するためには、オノマトペが含まれる大量の用例が必要となる。

香林らは、オノマトペの用例を日本語、英語、中国語、韓国語で表示するオンライン多言語辞書の開発を行った[2]。この研究は、小説から手動で用例を抽出する手法を用いているため、質の高い辞書を実現している。しかし、大量の用例を手で抽出（あるいは作成）するとコストが高い上、オノマトペの新しい用法を網羅することができない。

Asagaらは、オノマトペが用いられている文章をWebコーパスから自動抽出し、提示する辞典の開発を行った[3]。しかし、用例抽出手法が非常に単純で、Web上に存在する現代のオノマトペの用法を網羅的に抽出しているとはいえない。研究成果として一般に公開されているデータは80語余りのオノマトペを対象にしたもので、4,500語にも達するとされる日本語オノマトペと比較するとデータ量が不十分である。

このように、日本語オノマトペの大規模なコーパスは未だに整備されていない。

一方、多くの用例に触れるという学習方法には日本語母語話者が抱いている細かな感覚が伝わりにくいという問題点がある。先行研究では、オノマトペから感じる印象を、音象徴などを用いて評価したものがある[4, 5, 6]。

我々は、オノマトペが使用されたときの話し手、あるいは書き手の感情という側面からこの問題を解決しようとしている。したがって、大規模なオノマトペ用例文コーパスを構築し、コーパス中の各オノマトペに感情情報を付与することを目指している。

本論文では本研究の第一段階として、オノマトペ用

[†] A Method of Onomatopoeia Sentences Extraction from Blog Entries

Yuzu UCHIDA, Kenji ARAKI and Jun YONEYAMA

^{*1} 青山学院大学 理工学部電気電子工学科
College of Science and Engineering, Aoyama Gakuin University

^{*2} 北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

例文を自動的に抽出する手法を提案する。本手法はオノマトペの後続要素と係り先を詳しく分析した結果に基づくものである。このようなタスクでは、機械学習等の統計的手法を用いることも考えられる。統計的手法を用いる場合、大量の訓練データをあらかじめ用意しなければならないという問題点がある。一方、本研究の抽出手法では、Web上のブログ記事をコーパスとして用いるため、特定の言語資源を必要としないという利点がある。また、分析用のデータを作成する過程で手作業を要するが、その作業は軽微なものである。

本研究では、Webから収集したブログ記事からオノマトペ用例文を抽出する。大量のデータを低コストで入手できる点、現代日本語の用法を反映している点、新聞等のコーパスよりも感情情報が含まれやすい点でブログ記事は有用である。

2章では、オノマトペの後続単語に着目した特徴分析について述べる。3章では、2章の結果を踏まえてオノマトペ用例文の自動抽出手法を提案し、性能評価を行う。4章では、オノマトペの係り先に着目した特徴分析について述べる。5章では、4章から得られた知見に基づき、オノマトペ用例文の自動抽出手法を改善し、性能評価を行う。最後に、6章で本論文の結論と今後の課題について述べる。

2. オノマトペの後続単語分析

ブログ記事からオノマトペ用例文を大量に抽出する手法を構築するために、オノマトペが使用されている文の特徴を分析する。本章では、分析用のオノマトペ用例文の生成と、オノマトペの後続要素に着目した分析結果について述べる。

2.1 分析用データの生成

2.1.1 ブログ記事収集

Yahoo!ブログ検索Web APIを用いて、オノマトペ1語を検索クエリとしたときの検索結果上位20件分のスニペットを取得した(2010年6月20日時点)。検索クエリの対象としたオノマトペは、日本語オノマトペ辞典[7]に、動作を表すオノマトペとして掲載されている299語である。表1にその一部を示す。なお、使用した日本語オノマトペ辞典には、古事記などの古典から現代に至るまでの用例が掲載されており、見出し語は4,564語となっている。

この方法で、5,802件分のブログ記事に対するスニペットが得られた。本研究では、オノマトペを含む文とその近傍を分析対象とするため、各スニペットのTitle要素(ブログ記事のタイトル)とDescription要素(ブログ記事の本文)を使用する。したがって、合計

表1 収集対象としたオノマトペの語数と抜粋

動作カテゴリ	語数	オノマトペの例
騒ぐ	32	がやがや、どたばた
疲れる	13	うんざり、くたくた
働かない	15	ごろごろ、のんびり
吐く・もどす	11	がらがら、げろげろ
起きる・立つ	21	がぼつ、むっくり
飲む・酔う	67	がぶがぶ、ごくごく
食べる・かむ・舐める	71	かりかり、ばくり
見る・見える・睨む	53	ちらちら、まじまじ
咳をする・むせる	16	けほけほ、こほん

11,604の要素から重複(同一ブログ記事のスニペット)を除いた11,184要素を扱うことになる。

2.1.2 タグ付け

2.1.1項で収集したスニペットには正しくオノマトペが含まれていないものが多く存在する。例えば、「えへん(せきばらいの声を表すオノマトペ)」をクエリとして検索を行った結果、「今日はもう外に出えへんの。」という表現が取得される。オノマトペ用例文の自動抽出手法を考案するにあたり、正しいオノマトペ用例文と誤ったオノマトペ用例文を区別して分析しなければならない。そこで、取得したスニペットに人手で2種類のタグを付与した。タグ付けの作業者は日本語を母語とする20代の理系男子大学生1名である。

1つ目のタグは<omp>で、正しく用いられているオノマトペに付与される。2つ目のタグは<n-omp>で、オノマトペを表す文字列を含み、単語数が最小となる単語列に付与される。<n-omp>タグは、オノマトペ用例文抽出の際に、スニペット中に出現する誤りをストップワードとして用いる目的で付与する。汎用性の高いストップワード生成のために、単語数を最小にする。

オノマトペ“がぶり”をクエリとして得られたスニペットにタグを付与した実例を以下に挙げる。

- <omp>がぶり</omp>と嘯むと、口の中いっぱいに広がるじゅわじゅわとした甘酸っぱさが...
- ...先日から四十肩(五十肩?)<n-omp>がぶり返し</n-omp>てると思ってた店主だが...

2つ目の例で出現している“がぶり”という文字列は、格助詞の“が”と動詞の“ぶり返す”の一部であり、オノマトペの用例としては不適切である。したがって、<n-omp>タグ付与の対象となっている。

11,184要素にタグ付けを行ったところ、3,149の要素に<omp>タグが付与された。これにより、全体の28.2%の要素には正しくオノマトペが含まれていることが明らかになった。

2.1.3 文分割とフィルタリング

Description要素は複数の文から成り立っていることが多いため、1文単位に分割し、オノマトペが含まれる文のみを抽出する。文の区切り位置には、句点、疑問符、感嘆符、改行に加えて、ブログ記事の文末に多く見られる表現を用いる。具体的には、以下の表現である。

- ★, ☆
- ♪
- (笑), (笑), (汗), (汗)
- 2回以上連続した“w”あるいは“w”
- ^ ^, ^^
- …, …, …
- 鍵括弧

Title要素と文分割されたDescription要素のうち、オノマトペ用例文として不適切なものを除外する。ここでは、3つのヒューリスティクスを用いる。

- URLやメールアドレスが含まれる文を除外するため、半角英数字の連続(6字以上)をストップワードとする。
- 固有名詞として使用されているオノマトペを除外するため、鍵括弧で括られているオノマトペ、直後に敬称が続くオノマトペをストップワードとする。使用した敬称は、{さん, ちゃん, 君, くん, 様, さま, 殿}である。
- 名詞が含まれていない文、短い文(オノマトペ+4文字以下)は1文では意味を成さないと判断し、除外する。

2.1.2項で<omp>タグを付与されたTitle要素とDescription要素に対して上記の処理を施した結果、2,712のオノマトペ用例文が得られた。これらを正しいオノマトペ用例文として扱い、正解データと呼ぶ。

<omp>タグを付与されなかった——人手で誤りだと判断された——8,035の要素のうち、上記の処理を施した結果得られた1,160のオノマトペ用例文を不正解データと呼ぶ。単純なヒューリスティクスを用いたフィルタリングを行うだけで、誤った用例文を7分の1程度まで削減できることが明らかになった。

2.2 分析の手法と結果

2.2.1 形態素解析による後続単語の分析

オノマトペに後続する単語の特徴を分析するため、正解データ、不正解データに形態素解析を行う。形態

素解析器は日本語形態素解析システムJUMAN6.0[8]を用いた。形態素辞書はJUMANの附属辞書に本研究で使用する299語のオノマトペを副詞、普通名詞として加えたものである。

形態素解析の結果、オノマトペ、あるいはオノマトペに“と”が接続された語(“きっと”, “ぐっと”のように、オノマトペ+“と”が形態素辞書に1語として登録されている場合があるため)が1形態素として分割されたものは、正解データ中に1,706例、不正解データ中に212例存在した。それぞれについて、後続単語の品詞の内訳を図1、図2に示す。

動詞は正解データ中のオノマトペに後続する頻度が最も高く、全体の34.4%である。不正解データにおける割合(10.7%)の3倍を超えている。

図1 正解データにおけるオノマトペの後続単語の品詞(形態素解析でオノマトペが正しく分割された場合)

図2 不正解データにおけるオノマトペの後続単語の品詞(形態素解析でオノマトペが正しく分割された場合)

次に高い頻度で正解データ中のオノマトペに後続する品詞は助詞で、31.6%である。また、オノマトペ+助詞に後続する単語は、名詞、動詞が79.7%を占めている。不正解データの場合、助詞が後続する割合は全体の23.0%であり、そのうち47.3%が名詞、動詞に接続する。

3番目に高い頻度で正解データ中のオノマトペに後続するのは、句読点や括弧などの記号を表す“特殊”である。動詞、助詞と比較すると大きな割合ではないが、その内容に特徴がある。正解データでは句読点が180例中120例存在し、66.7%を占めている一方、不正解データでは46例中20例、43.5%である。

判定詞は正解データ、不正解データともに低頻度であるが、具体的な単語に相違点がある。正解データには“な”(判定詞“だ”の連体形)が20例存在したが、不正解データには1例も存在しなかった。

2.2.2 形態素解析で正しく分割されないオノマトペの後続単語の分析

2.2.1項で述べた形態素解析の結果、オノマトペ、あるいはオノマトペに“と”が接続された語が1形態素として分割されないものも多数存在した。具体的には、ABン型¹(“えへん”, “ごほん”, “とろん”), ABッ型(“きよろっ”, “じろっ”), ABリ型(“じろり”, “こくり”)等のオノマトペが正しく形態素解析されず、正解データ中の1,006例、不正解データ中の948例が該当した。

特定のオノマトペに関する用例文が分析対象にならないことは望ましくない。そこで、各文をオノマトペで区切り、オノマトペの直後の部分のみを形態素解析することで後続単語の分析を行う。正解データ、不正解データそれぞれについて、後続単語の品詞の内訳を図3、図4に示す。

正解データ中のオノマトペに後続する単語は、助詞が最も多い。後続する助詞には“と”, “が”, “っと”の3語があり、“と”が96.6%(282例)を占めている。不正解データでは、助詞が後続する割合は5.8%であり、低頻度である。

未定義語には疑問符や句読点、長音記号などが含まれ、正解データと不正解データで特に大きな差は見られなかった。

動詞は正解データ、不正解データで同程度の頻度だが、具体的な単語に相違点がある。

正解データでは、“する”(51例, 25.6%), “にる”(39例, 19.6%)の2語が突出して多く、“食べる”,

図3 正解データにおけるオノマトペの後続単語の品詞(形態素解析でオノマトペが正しく分割されなかった場合)

図4 不正解データにおけるオノマトペの後続単語の品詞(形態素解析でオノマトペが正しく分割されなかった場合)

“騒ぐ”, “飲む”などが続く。ただし, “にる”という動詞が高頻度になった背景には, 形態素解析誤りがある。オノマトペには助詞“に”が接続することが多くあるが, JUMANによる形態素解析結果では助詞“に”が動詞“にる”の連用形となっていた。これは, 用例文中のオノマトペの後ろに続く部分のみを切り出して形態素解析を行った結果, オノマトペの直後の助詞が文頭に置かれたことが原因である。

不正解データでは, “くる”(18例, 11.1%)が最も多く, “ぬく”, “ひる”などが続く。ただし, これらの動詞は“じーっ | くり / と”²や, “ぬっく | ぬく / した”, “ひっく | ひ / っ / く”といった表現から得られたものであり, 正解データと同様に形態素解析誤りが原因で高頻度になったものである。

1 日向[9]の方法に従い, オノマトペの語基をアルファベットで表記する。1つのアルファベットは1拍を表す。

2 | はオノマトペの境界, / は単語の境界を表す。

3. 後続単語分析に基づくオノマトペ用例文の自動抽出手法

3.1 アルゴリズム

2.2節で述べた用例文データの分析結果に基づき、オノマトペが正しく含まれる用例文を抽出する手法を提案する。本研究は実用に耐える用例文コーパスの構築を目標としているので、質の高いデータを自動的に収集する必要がある。したがって、再現率よりも適合率を優先させることを基本方針としている。

具体的なアルゴリズムを図5に示す。IV(h)にある“カテゴリ代表動詞”とは、オノマトペの動作カテゴリを明示する動詞を指し、第一著者が「見る、見つめる、見える、睨む、騒ぐ、働く、怠ける、起きる、立つ、疲れる、吐く、もどす、食べる、飲む、呑む、咳き込む、むせる」の17語と定義した。テストデータと

して2.1.1項で収集した分析用データに提案手法を適用すると、適合率1.000、再現率0.549となる。

3.2 評価実験

3.2.1 評価用データ

Yahoo!ブログ検索Web APIを用いて、オノマトペ1語を検索クエリとしたときの検索結果上位500件分のスニペットを取得した(2011年7月26日時点)。検索クエリの対象としたオノマトペは、2.1.1項で述べた299語と同じものとした。

この方法で、115,126件分のブログ記事に対応するスニペットが得られた。これらのスニペットのTitle要素とDescription要素、合計230,312の要素から重複を除くと225,446要素が残る。また、純粋なブラインドテストを行うため、アルゴリズムを構築する際に用いた分析用データとの重複データを削除する。以上より、分析用データと重複している1,122要素を除いた224,324要素を評価用データとし、図5のアルゴリズムを適用する。

3.2.2 結果と考察

まず、フィルタリング(IおよびII)の処理によって、87,184文が生成された。次に、JUMANを用いて形態素解析を行い、IIIの処理で31,960のオノマトペ用例文が抽出された。最後にIVの処理で17,774のオノマトペ用例文が抽出された。

抽出されたオノマトペ用例文から500文をランダムに選択し、人手で正否を判断したところ、466文が正しいオノマトペ用例文であった。この結果から、適合率93.2%が期待される。

次に、III、IVの各処理が本システムの性能に対してどの程度貢献しているのかを調査する。各処理で抽出されたオノマトペ用例文から200文ずつをランダムに選択し、人手で正否を判断したところ、処理IIIでは191文、処理IVでは184文が正しいオノマトペ用例文であった。適合率はそれぞれ95.5%、92.0%である。処理IVから得られた用例文は、オノマトペの直後のみを対象として正否の判定を行っているため、精度が劣っている。

実際に抽出されたオノマトペ用例文を以下に示す。

- ア 連日の猛暑にいささかうんざりしております。
 イ まわりのお嬢さんたちがみんな立ち上がって拍手をし、きゃーきゃー言っているのも、とてもよくわかる演技でありました。
 ウ 笛も一緒に煮込まれていて、これがこりっとした園ごたえのアクセントを生んでこれまた美味しい

図5 後続単語分析に基づくオノマトペ用例文抽出アルゴリズム

です。

エ ごぶごぶでやってた北新地のスイーツの店のパバ
ロアかなり美味そうやったなあ。

オ 喘息も少しででるけど、こんくらいやったら平気
でしょつ。

ア、イ、ウは正抽出例、エ、オは処理IIIで抽出さ
れた誤抽出例である

エは、“ごぶごぶ(飲み物を勢いよくのどに流し込ん
だ際の音などを表すオノマトペ)”の用例文として抽出
されたものであるが、同音の固有名詞(テレビ番組の
タイトル)が原因で誤抽出となっている。この誤抽出
は、オノマトペ直後の要素の表層的な情報を用いるだ
けでは回避することが困難であるが、オノマトペの係
り先を考慮することで、ある程度改善可能であると思
えられる。この例で“ごぶごぶ”の係り先である“やっ
てた”は、オノマトペ“ごぶごぶ”の係り先としては不
適切である。

オは“こん(軽く咳をする音などを表すオノマトペ)”
の用例文として抽出されたものであるが、“このくら
い”の口語表現の一部が誤って抽出されている。この
ような誤抽出は、“こん”、“かつ”、“くい”等の2文字
からなる短いオノマトペに顕著である。短いオノマト
ペをクエリとしたスニペットでは、クエリが別の単語
の一部に一致する可能性が高い。そのようなスニペッ
トに形態素解析を施すと、オノマトペ(正確にはオノ
マトペに一致する文字列)が名詞や動詞として認識さ
れやすいという特徴がある。

以上の分析結果より、形態素解析で得られたオノマ
トペ部分の品詞と、オノマトペの係り先を考慮して、
オノマトペ用例文抽出アルゴリズムに変更を加える。

4. 構文解析によるオノマトペの係り先分析

3章で提案したオノマトペ用例文抽出アルゴリズム
を改善するために、オノマトペの係り先に着目した分
析を行う。

表2 正解データ中のオノマトペの係り先(正係り先)

動作カテゴリ	オノマトペの係り先
騒ぐ	する, 言う, 騒ぐ
疲れる	する, 疲れる, 痩せる
働かない	する, 過ごす, 転がる
吐く・もどす	する, 吐く, やる
起きる・立つ	立つ, 起き上がる, 入る
飲む・酔う	飲む, する, 食べる
食べる・かむ・舐める	する, 食べる, なる
見る・見える・覗む	する, 見る, 眺める
咳をする・むせる	言う, する, 咳き込む

4.1 正解データの構文解析結果

正解データを形態素解析した結果、オノマトペが1
形態素として分割された1,706文を対象に構文解析を
行った。構文解析には、日本語構文解析システム
KNP10)を使用した。各動作カテゴリに含まれるオノ
マトペの係り先のうち、出現頻度の高い3語を抜粋し
て表2にまとめる。

4.2 評価用データから抽出されたオノマトペ用例文の 構文解析結果

3.2.1項で述べた評価用データのフィルタリング処
理後のデータを形態素解析し、オノマトペが1形態素
として分割された41,724文を対象に構文解析を行っ
た。

オノマトペに係る単語の品詞は動詞が最も多く、全
体の72.0%を占めている。次いで、名詞(18.9%), 形
容詞(5.6%)となる。ここではオノマトペに係る動詞
に限定して分析を行う。

評価用データにおけるオノマトペの係り先と正係り
先との一致率を表3に示す。ここで、“正係り先”とは
4.1節で述べた正解データ中のオノマトペの係り先を
指す。評価用データにはオノマトペの係り先である動
詞は延べ30,039語あり、そのうち15,630語が正係り先
と一致している。また、一致したデータからランダム
に選んだ200文を人手で調査したところ、199文が正
しいオノマトペの用例文であった。誤りだと判断され
た用例文は、“(ゼーゼー)”ということで、タイミング
をどう読むか考え中です。”で、

「混雑でサウロタッチ(あるかどうか知らないけれ
ど)ができなかったら、その日は眠れない。(ゼー
ゼー)

ということで、タイミングをどう読むか考え中
です。」

というブログ記事から抽出されたものである。“(ゼー

表3 正解データと自動抽出されたオノマトペ用例文
の係り先の一致率

動作カテゴリ	係り先一致率
騒ぐ	52.7%
疲れる	61.5%
働かない	58.6%
吐く・もどす	57.6%
起きる・立つ	57.4%
飲む・酔う	44.6%
食べる・かむ・舐める	45.7%
見る・見える・覗む	62.3%
咳をする・むせる	40.3%

ぜー)”は直前の文を修飾しており、後ろの文との関連はない。これは、句点を文の境界として文分割を行ったことによる誤りである。

以上の結果より、オノマトペの係り先に基づいた用例文抽出手法の有効性が示唆された。

5. 後続単語及び係り先分析に基づくオノマトペ用例文の自動抽出手法

5.1 アルゴリズム

構文解析による用例文データの分析結果に基づき、3.1節で述べたオノマトペ用例文の抽出手法に改良を加える。具体的なアルゴリズムを図6に示す。V'(a)、V'(d)にある“オノマトペ(副詞)”とは、形態素解析によってオノマトペが副詞として認識されていることを示し、V'(b)、V'(e)にある“オノマトペ(副詞以外)”とは、形態素解析によってオノマトペが副詞以外として認識されていることを示す。

2.1.1項で収集した分析用データをテストデータとして提案手法を適用すると、適合率1.000、再現率0.642となる。

5.2 評価実験

5.2.1 評価用データ

Yahoo!ブログ検索Web APIを用いて、オノマトペ1語を検索クエリとしたときの検索結果上位500件分のスニペットを取得した(2012年1月24日時点)。ただし、分析に用いたデータとの重複を避けるため、1か月以内に更新された記事に限定した検索結果を用いる。検索クエリの対象としたオノマトペは、2.1.1項で述べた299語と同じものとした。

この方法で、41,315件分のブログ記事に対応するスニペットが得られた。これらのスニペットのTitle要素とDescription要素、合計82,630の要素から重複を除くと80,600要素が残る。さらに、アルゴリズムを構築する際に用いた分析用データと重複しているデータを削除した80,349要素を評価用データとし、図6のアルゴリズムを適用する。

5.2.2 結果と考察

まず、I'、II'のフィルタリング処理によって、27,710文が生成された。次に、IV'の処理によって、オノマトペの係り先を手がかりにして、5,574文のオノマトペ用例文が抽出された。さらに、オノマトペの後続要素を手がかりにして、V'の処理から6,293文、VI'の処理から3,570文のオノマトペ用例文が抽出された。

抽出されたオノマトペ用例文から500文をランダムに選択し、人手で正否を判断したところ、481文が正

しいオノマトペ用例文であった。この結果から、全体の適合率は96.2%が期待される。アルゴリズムの変更によって適合率が3.0ポイント上昇し、本手法の有効性が示された。

図7に、動作カテゴリごとの各処理から得られたオ

- I' 収集したスニペットに対し、2.1.3項で述べた手法で文分割及びフィルタリングを行う。
- II' 2.1.2項で述べた<n-omp>タグが付与された文字列をストップワードとしてフィルタリングを行う。
- III' 各文に形態素解析を行い、オノマトペが1形態素として分割された場合、構文解析を行う。
- IV' III'でオノマトペの係り先が正係り先と一致した場合、適切なオノマトペ用例文と判断する。
- V' III'でオノマトペの係り先が正係り先と一致しなかった場合、以下の条件を満たす文を適切なオノマトペ用例文と判断する。
 - (a) オノマトペ(副詞) + 助詞 + 動詞
 - (b) オノマトペ(副詞以外) + 助詞 + 正係り先の動詞
 - (c) オノマトペ + 助詞 + 名詞
 - (d) オノマトペ(副詞) + 動詞
 - (e) オノマトペ(副詞以外) + 正係り先の動詞
 - (f) オノマトペ + 判定詞
 - (g) オノマトペ + 句読点
- VI' III'でオノマトペが1形態素として分割されなかった場合、以下の条件を満たす文を適切なオノマトペ用例文と判断する。
 - (h) オノマトペ + 助詞 “と”
 - (i) オノマトペ + 動詞 “する”
 - (j) オノマトペ + カテゴリ代表動詞
 - (k) オノマトペ + “に”

図6 後続単語及び係り先分析に基づくオノマトペ用例文抽出アルゴリズム

図7 各処理で自動抽出したオノマトペ用例文数

ノマトペ用例文の数を示す。オノマトペカテゴリによって、各処理で得られる用例文数の割合が大きく異なることがわかる。特に、“起きる”と“見る”については処理VI'に大きな特徴がある。

本研究では45語のABッ型オノマトペを用いているが、それらすべてが形態素解析によって1語として認識されなかった。このように、ABッ型オノマトペは形態素解析によって1語として認識されない傾向が強い。“起きる”カテゴリには、“がぱっ”、“すくっ”、“すたっ”、“ぼさっ”、“むくっ”等、ABッ型のオノマトペが多いため、処理VI'によって多くの用例文が抽出された。一方、“見る”カテゴリでは、実験データ6,431文のうち81.5%にあたる5,244文が形態素解析に成功している。したがって、処理VI'による用例文抽出の割合が4.3%と非常に小さくなった。

最後に、IV', V', VI'の各処理が本システムの性能に対してどの程度貢献しているのかを考察する。各処理で抽出されたオノマトペ用例文から200文ずつをランダムに選択し、人手で正否を判断したところ、処理IV'では197文、処理V'では196文、処理VI'では172文が正しいオノマトペ用例文であった。適合率はそれぞれ98.5%、98.0%、86.0%である。処理VI'から得られた用例文は、オノマトペの直後のみを対象として正否の判定を行っているため、他の処理よりも精度が大きく劣っている。

処理VI'で誤抽出された用例文は、“わっ”を対象としたものが11文、“ぐい”が5文、“すっ”が4文、“かりん”が3文、“くい”、“ぐー”、“げー”、“しゃん”、“はた”がそれぞれ1文となっていた。

“わっ”に関する誤抽出の中には、“ぶわっ”と噴き出す”、“ぞわっとする”、“ふわっ”と香る”など、他のオノマトペの語末によるものが7文存在した。このような誤抽出は、本手法では対象としていなかったオノマトペもあらかじめ辞書として保持しておくことで、解

決できると考えられる。

“ぐい”に関する誤抽出は全て“ぐい飲み(呑み)”という名詞が原因となっており、VI'(j)で抽出されている。このような誤抽出を防ぐには、“ぐい”は直後に助詞を伴って使用されるというヒューリスティクスが有効だと考えられる。

“すっ”と“わっ”に関する誤抽出の一部はブログ記事に特徴的に表れる口語表現に原因がある。具体的には、“すっ”では“そうすっ”や“マジレスすっ”のように“すると”の促音便が、“わっ”では“もういいわっ”と、逃げようとする”のような表現が原因となっている。これらの誤抽出は文字数の少ないAッ型に起こりやすいと考えられる。また、後続単語の品詞などの情報では抑制することが困難だと予想されるため、カテゴリ代表動詞のように後続する単語を限定する手法を考案する必要がある。

6. おわりに

本論文では、日本語オノマトペの大規模なコーパス構築を目指し、ブログ記事を対象としたオノマトペ用例の自動抽出手法を提案した。提案手法は、オノマトペの後続要素と係り先の分析から明らかになった、オノマトペを含む文の特徴を利用したものである。299語のオノマトペをクエリとして得た41,315件分のブログ記事に本手法を適用した結果、15,437文の用例文が抽出された。抽出の適合率は96.2%であり、大量の文書集合や語彙資源を必要としないシンプルな手法であるが、良好な結果となった。

今後は、抽出した用例文をオノマトペの意味によって分類し、さらに感情情報を付与したコーパスを構築する予定である。

謝辞

本研究は科研費(課題番号:23700256)の助成を受けたものである。

参考文献

- [1] 三上京子:日本語教育のための基本オノマトペの選定とその教材化, ICU 日本語教育研究, Vol.3, pp.49-63, 2006.
- [2] 香林隆子, 増永良文:オノマトペのオンライン多言語辞書の構築, DEWS2002 論文集, A4-4, 2002.
- [3] Chisato Asaga, Mukarramah Yusuf and Chiemi Watanabe: Onomatopedia: Onomatopoeia Online Example Dictionary System Extracted from Data on the Web, The 10th Asia Pacific Web Conference (APWeb), 2008.
- [4] 市岡健一, 福本文代: Web 上から取得した共起頻度と音象徴によるオノマトペの自動分類, 電子情報通信学

- 会論文誌, D, 情報・システム J92-D (3), pp.428-438, 2009.
- [5] 小松孝徳, 清河幸子, 秋山広美: オノマトペか感じる印象を表現する属性とその客観的数値化, HAI シンポジウム 2009, 2B-4, 2009.
- [6] 中部文子, 浅賀千里, 渡辺知恵美: 感性情報を利用したオノマトペ学習システムの開発, 第1回データ工学と情報マネジメントに関するフォーラム (DEIM2009), E5-1, 2009.
- [7] 小野正弘: 擬音語・擬態語 4500 日本語オノマトペ辞典, 小学館, 2007.
- [8] 日本語形態素解析システム JUMAN 6.0, <http://www-lab25.kuee.kyoto-u.ac.jp/nlresource/juman.html>.
- [9] 日向茂男: 語形からみた擬音語・擬態語, 東京学芸大学紀要. 第2部門, 人文科学 42, pp.59-70, 1991.
- [10] 黒橋禎夫, 長尾眞: 並列構造の検出に基づく長い日本語文の構文解析, 自然言語処理, Vol.1, No.1, pp.35-57, 1994.
- (2011年10月23日 受付)
(2012年4月5日 採録)
- [問い合わせ先]
〒252-5258 神奈川県相模原市中央区淵野辺5-10-1
青山学院大学 理工学部電気電子工学科
内田 ゆず
TEL: 042-759-6418
E-mail: yuzu@ee.aoyama.ac.jp

著者紹介

うちだ ゆず [正会員]

2005年北海道大学工学部情報工学科卒業。2010年同大学院博士後期課程修了。博士(情報科学)。同年, 青山学院大学理工学部助手。2012年より同助教, 現在に至る。自然言語処理, 特に概念獲得システム, 日本語オノマトペなどに関する研究に従事, 日本知能情報ファジィ学会, 電子情報通信学会, 各会員。

あらき けんじ [正会員]

1982年北海道大学工学部電子工学科卒業。1988年同大学院工学研究科電子工学専攻博士後期課程修了。同年, 北海学園大学工学部, 電子情報工学科助手。1989年同講師, 1991年同助教, 1998年同教授。1992-1993年, 2000年, 2009年米国スタンフォード大学CSLI客員研究員。1998年北海道大学大学院工学研究科電子情報工学専攻助教授。2002年同教授を経て2004年北海道大学大学院情報科学研究科メディアネットワーク専攻教授, 現在に至る。工学博士。自然言語処理, 特に対話処理, 機械翻訳などの研究に従事。電子情報通信学会, 情報処理学会, 人工知能学会, 言語処理学会, 日本認知科学会, ACL, IEEE, AAAI 各会員。

よねやま じゅん [正会員]

1996年University of California, Los Angeles電気工学科博士後期課程修了(Ph.D.)。同年静岡大学工学部電気電子工学科助手, 2000年青山学院大学理工学部電気電子工学科助教授, 2007年同准教授。現在, 同教授。ファジィシステム, むだ時間システム, ロバスト制御, 確率システム制御, サンプル値システム制御などの研究に従事。日本知能情報ファジィ学会, IEEE, 計測自動制御学会, 電気学会, 日本ロボット学会などの各会員。

A Method of Onomatopoeia Sentences Extraction from Blog Entries

by

Yuzu UCHIDA, Kenji ARAKI and Jun YONEYAMA

Abstract :

There are numerous onomatopoeic words unique to the Japanese language. Those words are indispensable to make smoother communication in Japanese. Though native speakers of Japanese normally use onomatopoeias without thinking, they can hardly restate them in different terms, let alone explain what they mean. It is said that onomatopoeias are quite hard to master for people studying Japanese. One way to do this is the exposure to many sentences that contain onomatopoeia, and hence we propose an automatic onomatopoeia sentences extraction method based on an analysis of example sentences. As a result, our method extracts 15,437 onomatopoeia sentences from 41,315 blog entries and achieves an accuracy of 96.2%(481/500).

Keywords : onomatopoeia, subsequent word, syntax analysis, corpus

Contact Address : **Yuzu UCHIDA**

College of Science and Engineering, Aoyama Gakuin University

5-10-1, Fuchinobe, Tyuo-ku, Sagami-hara-shi, Kanagawa, 252-5258, JAPAN

TEL : 042-759-6418

E-mail : yuzu@ee.aoyama.ac.jp