5T-05

テキストマイニングシステム STM におけるオノマトペ分析

小野　祥太郎[†]　　谷津　元樹[‡]　　原田　実[‡]
青山学院大学理工学部情報テクノロジー学科[†]

1. はじめに

近年、SNS などの発達により、消費者が感覚的な意見を発信する機会が多くなっている。擬音語・擬態語、つまりオノマトペは元来、商品やサービスに対する消費者の直感的な意見が含まれていると考えられてきた。インターネット上投稿されたアンケートなどに含まれるオノマトペを調査することで、商品やサービスの改善点の模索等に役立つと考え、原田研究室で研究されてきたテキストマイニングシステム STM[1]に、オノマトペを抽出し、他の選択回答部分との相関関係を分析する機能を追加した。

2. 従来のオノマトペ抽出の問題点

今までのオノマトペ研究でも、オノマトペを抽出し様々な分析を行う研究は数多く行われてきた。しかし、オノマトペを抽出する段階で、辞書登録されているオノマトペが少なかったり、解析間違いしやすいオノマトペは対象から除外したり、同音異義のオノマトペを同じように扱ったり、と曖昧で網羅的でない抽出が多かった。本研究では、オノマトペがもともと多く登録されている EDR 電子化辞書を使用し、そこに内田らの研究[2]で使用されたオノマトペを新たに登録、擬音語・擬態語 4500 日本語オノマトペ辞典[3]を参考に、カテゴリ分けによる意味分類を行うことでこれらの問題を解決する。

3. オノマトペに関する辞書体系

EDR 電子化辞書は、数多くの概念と、その概念間の上位-下位関係、概念を語意として持つ単語で構成されている。本研究以前は、オノマトペは「擬音語や擬態語で表される様態」の下位概念にまとめられていたが、意味分類などはされておらず、この概念の下位以外の場所にもオノマトペを含む概念が EDR 電子化辞書内に散らばっていた。そのため、内田らの研究で用いられた 647 語のオノマトペを新たに辞書登録する際に、「擬音語や擬態語で表される様態」の下位概念として「自然」、「人間」、「事物」、「その他」の 4 つの大カテゴリを作成し、さらに、「自然」の下には「天気」、「温度」、「水・液体」、「火・土」を、「人間」の下には、「動作・状態」、「感情・感覚」、「性格・性質」、「体格・姿」を、「事物」の下には「動き・変化」、「形・状態」、「音・道具・金銭」、「程度」を小カテゴリとして作成し、これらの適切な小カテゴリの下にこの 647 語を配置した。その後、もともと登録されていたオノマトペの概念と、新たに配置したオノマトペと同音かつ「擬音語や擬態語で表される様態」の下位概念以外の場所にあるオノマトペをカテゴリ分けし、各カテゴリの下位概念として配置した。この概念の体系を図 1 に示す。

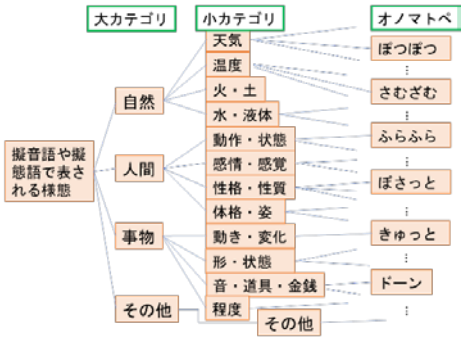

図 1 オノマトペの概念体系

4. オノマトペ抽出のアルゴリズム

テキストマイニングシステム STM は、意味解析システム SAGE[4]に自由記述文を送ることによって得られた意味グラフを利用している。本研究のオノマトペ検出方法は、この意味グラフから副詞で、かつ上位概念に「擬音語や擬態語で示される様態」をもつ形態素を抽出し、抽出できた場合、この文をオノマトペが存在する文とし、抽出したオノマトペを検出できたオノマトペとして扱う。そして、上位概念を遡る際に、途中に出てきた、大カテゴリ、小カテゴリを意味分類として利用する。この意味割り当ての例を図 2 に示す。

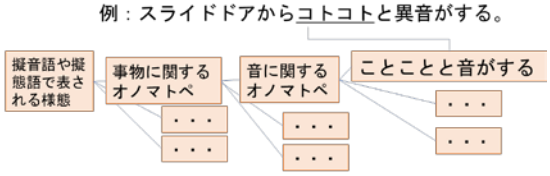

図 2 オノマトペの意味割り当ての例

5. オノマトペの分析機能

5.1. 頻度分析

検出されたオノマトペを図 3 のように棒グラフや円グラフなどで表す機能を実装した。

Onomatopoeia analysis in semantic text mining system STM

Ono Shotaro† Yatsu Motoki‡ Harada Minoru‡
†Faculty of Science and Engineering, Department of Integrated Information Technology, Aoyama Gakuin University.

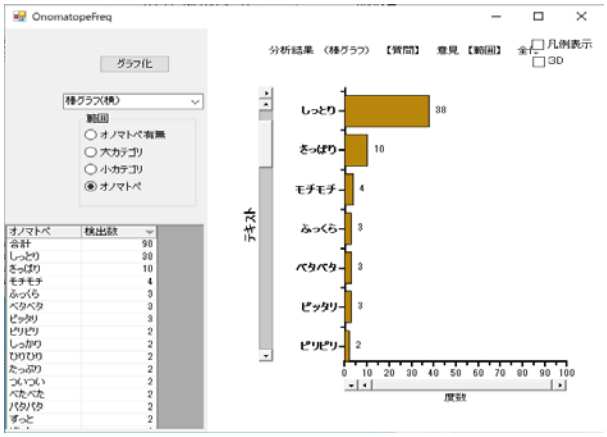

図 3 頻度分析機能

5.2. 属性分析

検出したオノマトペの、回答者の年齢といった属性値ごとの頻度を、図 4 のように表す機能を実装した。

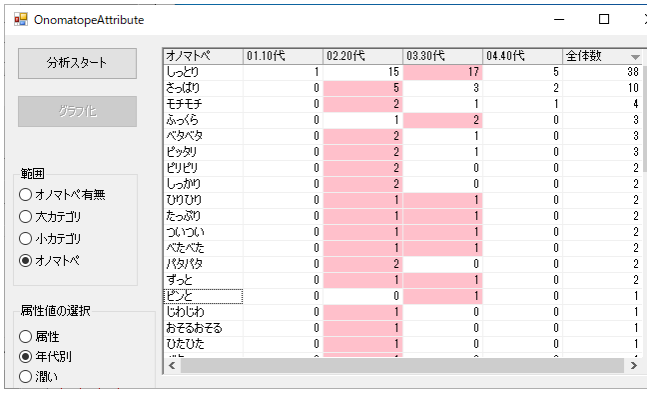

図 4 属性分析機能

6. 評価実験

評価実験として、国土交通省が公開している自動車不具合データを用いる。このデータのうち、某自動車メーカーの 2006 年から 2018 年までの不具合情報 6420 件を対象にテキストマイニングを行った。また、2010 年の那須川らの研究[5]による結果と比較を行った。

7. 実験結果

評価実験の結果、6420 件のデータのうち、オノマトペが検出されたのは、311 件だった。頻度分析の結果、最も多かった大カテゴリ、小カテゴリ、オノマトペはそれぞれ、「事物」で 248 件、「音・道具・金銭」で 108 件、「ガラガラ」で 55 件だった。属性分析の結果は、図 5 のようになった。

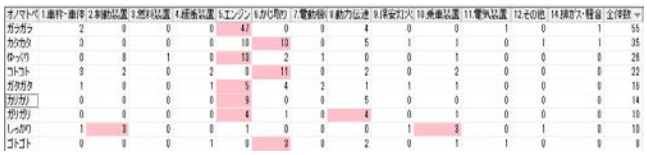

図 7-1 評価実験の属性分析の結果

8. 評価と考察

那須川らの研究では、自動車不具合データから擬音語をパターンによって抽出し、それを辞書登録したうえで辞書ベースの検出を行っている。その結果、24458 件のうち 521 件の擬音語が検出されている。本研究と那須川らの研究との、頻度順における上位 9 件のオノマトペの比較を図 6 に示す。

オノマトペ	検出数
ガラガラ	55
カタカタ	35
ゆっくり	26
コトコト	22
ガタガタ	16
カリカリ	14
ガリガリ	10
しっかり	9
ゴトゴト	8

キーワード	頻度
ガタガタ	93
カタカタ	44
カラカラ	29
ガラガラ	29
キーキー	22
ガリガリ	19
ギシギシ	17
ゴトゴト	14
ガクン	13

図 6 頻度分析比較(左:本研究 右:那須川らの研究)

図 6 に示されている通り、順位こそ違うが、多くの擬音語が二つの結果で共通していることがわかる。また、不具合装置との相関を比較した際も、動力伝達の不具合と「ガクン」、制動装置の不具合と「キーキー」の相関が強いといった、那須川らの研究と同一の発見をすることができた。最終的には、本研究により、那須川らの研究のようなパターン抽出や、辞書登録といった作業を介さない、自動化されたオノマトペ分析機能を実装することができたといえる。今後の改良としては、新語登録や EDR 辞書内に散らばるオノマトペの上位下位関係の整備、辞書内の共起表現の頻度や単語の頻度を調整することによる精度向上が考えられる。

参考文献

[1] 西脇　剛,保立哲志,原田実: “意味解析に基づくテキストマイニングシステム STM”,情報処理学会第 69 回全国大会論文集,2C-03,第 2 分冊 pp. 89-90. (2007.3).

[2] 内田ゆず,荒木健治:“クラスタ分析を用いた商品レビューに含まれるオノマトペに基づく商品カテゴリの類型化”,人工知能学会論文誌,30 巻,第 1 号,pp246-256.(2015.1).

[3] 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.

[4] 原田実,水野 高宏: “EDR を用いた日本語意味解析システム SAGE ", 人工知能学会論文誌, Vol.16, No.1, pp.85-93.(2001.1).

[5] 那須川哲也,海野裕也,村上朋子:“機器の不具合を記述した日本語と英語のコーパスにおけるオノマトペ”,言語処理学会第 16 回年次大会発表論文集,pp154-157.(2010.3).