E-035

# 出現頻度と周辺文脈に基づくツイート中の新語の検出
# Neologism Detection from Tweets Based on Frequency of Appearance and Context

向井 勇希† 杉本 徹†
Yuki Mukai Toru Sugimoto

## 1. 研究背景と目的

現代社会において、私たちは世相や流行に応じて新しい言葉を生み出し使用している。しかし、こうした言葉は形態素解析を必要とするシステムにおいて正しく処理できないことが多い。これを解決するためには新語の辞書への追加が必要となるが、新語は日々新たなものが生み出されており、その多様性も相まって正確かつ継続的に追加することは困難である。

先行研究として、大規模なウェブアーカイブから新語を抽出するものがある[1]。この研究では文脈情報を用いて抽出した新語に対して、出現頻度の時系列的変化および格交替などの認知言語学の観点から分析を行っている。別の関連研究として、Twitter 上の投稿（以降、「ツイート」と表記）から俗語として用いられている単語を検出するものがある[2]。しかし、対象となる単語は俗語以外の用法が既に辞書に登録されている単語であり、未知の単語を検出することはできない。

本研究では、ツイートから N-gram の集合を抽出し、出現頻度の高いものを新語の候補として抽出する。その上で、出現確率の時間的変化や出現文脈の特徴によって候補を絞り込むことで辞書へ追加する新語の候補を効率的に検出することを目指す。

## 2. 研究の概要

### 2.1 システム概要

図 1 にシステムの処理の流れを示す。

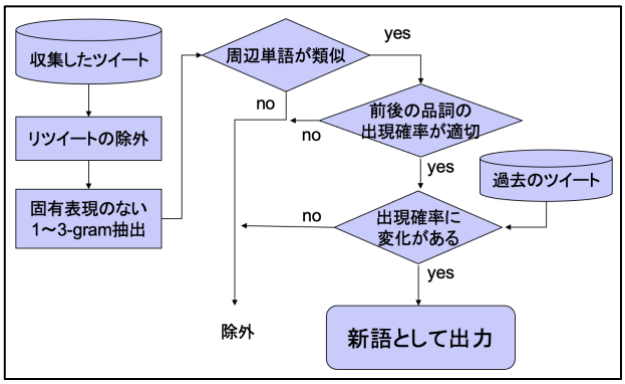


図 1　システムの処理の流れ

日本語のツイートを収集し、その中からリツイートや画像の URL といった不要な要素をテキストから除外する。この処理を施したツイートから N-gram（N=1, 2, 3）を抽出し、出現確率が閾値以上の N-gram を新語の候補とする。続いて周辺単語の類似度、前後の品詞、および出現確率の時間的変化によって候補の絞り込みを行い残ったものを新語として出力する。

† 芝浦工業大学　Shibaura Institute of Technology

### 2.2 新語の定義と対象コーパス

新語とは、新たに生まれた事物や概念を表現するために用いられる単語を指す。本研究では「一つの単語としては既存の辞書には未登録かつ固有表現でない」単語を新語として定義する。こうした単語は話し言葉や SNS に多く見られるため[3]、ツイートを研究の対象とした。

## 3. 研究の結果

### 3.1 ツイートの収集

2020 年 12 月 28 日に投稿された日本語のツイートを 120,000 件取得し、新語の抽出を目指す対象とした。ツイートに含まれる各単語の出現頻度が後の処理に影響を及ぼすと考えられるため、1 時間ごとに 5,000 件のツイートを取得し、時間帯による投稿内容の偏りが生じないデータセットを作成した。

### 3.2 N-gram の抽出

予備実験の結果、新語の形態素解析では誤って複数の形態素に分割しているものが多く、2 または 3 形態素に過分割されているものがほとんどであった。このことから本研究では固有表現および URL などの記号列を含まない 1-gram〜3-gram を抽出し、その中で出現確率が閾値以上のものを新語の候補とする。

本研究では、手法の検討と閾値の決定のためのサンプルデータとして、抽出した 1-gram〜3-gram から新語と呼べる N-gram とそうではない N-gram を 10 個ずつ選んだ。表 1 に選んだ N-gram を示す。

### 3.3 周辺文脈の類似度を用いた N-gram の絞り込み

新語の候補とした N-gram が明確に定まった意味を持つ形態素列であるならば、その周辺文脈に何らかの特徴があるはずである。そこで、周辺文脈の類似度を用いて N-gram の絞り込みを行う。対象の N-gram を含むツイートを自然言語処理ライブラリ spaCy の学習済みモデルを用いてベクトル化し、同じ対象を含むツイート間でそれぞれのベクトルのコサイン類似度を計算する。こうして得られた類似度の平均が一定の閾値を下回った N-gram を新語の候補から除外する。

### 3.4 前後の品詞の分布を用いた絞り込み

事前に収集したツイートから名詞と動詞の前後に出現する品詞の出現確率の分布$P_{noun}, P_{verb}$を計算する。次に対象 N-gram を含むツイートを検索し、そこから N-gram の前後の品詞の分布$Q$を計算する。ある N-gram が文中において名詞として用いられているならば分布$P_{noun}$と$Q$が、動詞であるならば$P_{verb}$と$Q$が似たものになると仮定し、分布間のカルバック・ライブラー情報量$D_{KL}(P_{noun}\|Q)$および$D_{KL}(P_{verb}\|Q)$が共に一定の閾値以上となる N-gram を除外する。

表１　手法の検討に用いた N-gram

| 新語 | N-gram | 新語でない |
|---|---|---|
| リプ　ガチ<br>フォロバ　キャス<br>ぴえん　タメ<br>フリート　コピペ<br>イケボ　ヲタ | 1-gram | ます　ください<br>から　てる　お願い<br>出来　参戦　好き<br>自分　今年 |
| おつ　リアタイ<br>コロナ禍<br>電子書籍　人狼<br>固ツイ　本垢<br>ワンチャン<br>リア友<br>ハッシュタグ | 2-gram | して　ました<br>者募集<br>ございます　の匿名<br>た人<br>おはようござい<br>よろしくお願い<br>仕事納め　頑張って |
| うぽつ　おめー<br>寝落ち（し）<br>おはあり<br>おつあり<br>ぶっちゃけ<br>二次創作<br>もふもふ<br>おはおは<br>おっけー | 3-gram | 参加者募集<br>ありがとうございます<br>おはようございます<br>お願いします<br>んだけど　んだよ<br>疲れ様です<br>なんだ　せて頂き<br>をお過ごしの |

### 3.5 出現確率の時間的変化に基づく絞り込み

「新語は時間の経過とともに使用される頻度が増す」という仮説を立てる。この仮説に該当するかを判定するために、それぞれの N-gram について以下の手順で投稿時期と出現頻度の相関を計算する。

① 対象 N-gram を含むツイートを現在から新しい順に 1000 件収集し、その中で最も新しいものと古いものの投稿時刻の差を記録する。
② 遡って収集する起点を 3 ヶ月前に変更し、同様に投稿時刻の差を記録する。
③ 起点を 3 ヶ月ずつ変更しながら 3 年前のツイートに到達するまで①と②を繰り返す。

記録した時刻の差の最大値と最小値に 10 倍以上の開きがある場合を投稿時期と出現頻度に相関があるものとして絞り込みを行う。

表１の N-gram を用いてそれぞれの閾値を調整した結果、表 1 の全 N-gram の中で新語の候補として最終的に残った N-gram を表２に示す。

表２　絞り込みの結果残った N-gram

| N-gram | 新語候補 |
|---|---|
| 1-gram | リプ　ぴえん　フリート　コピペ<br>イケボ　好き |
| 2-gram | リアタイ　人狼　固ツイ　ワンチャン<br>本垢　リア友　仕事納め　頑張って |
| 3-gram | おはあり　おつあり　おめー　二次創作<br>もふもふ　おはおは　おっけー<br>ありがとうございます<br>おはようございます　お願いします |

### 3.6 評価実験

以上に挙げた絞り込みの方法及び実験を通して得られた適切な閾値を以下の手順でデータセットに適用し、システムの性能の評価を行った。

① データセットからツイートを無作為に 1000 件選択する。
② 選択したツイートに含まれる新語を人手で列挙する。
③ 選択したツイートに対してシステムを適用し、新語を抽出する。
④ ②と③の結果がどれほど一致しているか比較する。

表３に正しく抽出できた新語の例を示す。

表３　正しく抽出できた新語の例

| N-gram | システムで抽出 |
|---|---|
| 1-gram | 推し　コロナ　ガチャ　コメ　ヲタ<br>アバター　ニチャァ　匂わせ |
| 2-gram | リア友　クラメン　クラスター<br>不思議ちゃん　婚活　電子書籍 |
| 3-gram | 声真似主　グリーン成長戦略<br>かます　第三波　二次創作　おめー |

人手で列挙した N-gram のうち、システムが正しく抽出できたものの割合は、1-gram が 43%、2-gram が 56%、3-gram が 40%であった。また、絞り込みで最後まで残った N-gram のうち、人手で列挙した N-gram にも含まれるものの割合は 1-gram が 41%、2-gram が 47%、3-gram が 46%であった。

## 4. 考察

3 種類の絞り込みそれぞれによって除かれた N-gram の内容と量を確認した。その結果、絞り込みの前後を比較した N-gram の除去率が最も低かったのは周辺文脈の類似度による絞り込みであった。これは絞り込み意味が定まった N-gram でも周辺文脈の類似度が低いものが多く、閾値を低く設定せざるを得なかったことが原因である。

逆に除去率が最も高かったのは前後の品詞の出現確率による絞り込みであった。これは周辺文脈の類似度による絞り込みを目指したような、定まった意味を持たない N-gram も合わせて除外されたことが理由であると考えられる。

## 5. 結論

ツイートから N-gram のリストを生成し、複数の条件で絞り込みを行うことで新語の抽出を行った。その結果、リストに含まれる新語のうち 5 割弱を抽出することができた。しかし、新語ではない N-gram に絞り込みで除き切れないものが多く、絞り込み方法の見直しや新たな条件の追加によって改良する余地も大きい。また、本研究ではツイートから新語を抽出する部分を扱ったが、辞書の追加などに応用するためには品詞や意味内容の情報も合わせて必要である。抽出した新語にこれらの情報を付与することが実用性を高める上での課題である。

### 参考文献

[1] 鍛治 伸裕, 宇野 良子, 喜連川 優, “言語学研究の支援を目的とした大規模時系列ウェブアーカイブからの新造語のマイニング”, DEIM フォーラム 2009 発表論文集, C6-2 (2009).
[2] 青木 竜哉, 笹野 遼平, 高村 大也, “ソーシャルメディアにおける俗語の検出”, 言語処理学会第 23 回年次大会発表論文集, pp.322-325 (2017).
[3] 黎 斯琳, “現代日本語の若者語の研究”, 広島大学日本語・日本文化研修プログラム研修レポート集, 22 期巻, pp.42-55 (2008).