MeCab の辞書強化とツイート分析

鹿志村 萌生　(指導教員：粕川 正充)

1　はじめに

オープンソースの形態素解析エンジンである MeCab には主として IPA 辞書が使われているが,IPA 辞書は固有名詞に弱い. また, 近年ネットスラングや 2 つ以上の意味を持つ新しい言葉が増えてきており,IPA 辞書だけでは対応するのが難しい. そこで,MeCab の辞書自体を強化することによって, 解析結果が “未知語” となってしまう語句を解析できるようにし,Twitter において, 指定した単語を含むツイートを分析することとした.

2　ツイートの取得

プログラミング言語 Python[1] を使って Twitter の API を操作した.Twitter API Key を取得し,Python のライブラリである Tweepy を用いた.Tweepy では, 自分の Twitter アカウントと絡めてタイムライン上の最新ツイートやフォロー数, フォロワー数などの情報を取得したり,Twitter を開かなくてもつぶやきを更新したりすることができる.
取得範囲は Twitter の全ユーザを対象とした. ツイートの本文に指定の単語を含むものを最新のものから順に取得していく.

3　MeCab の辞書

3.1　辞書の強化

今回 MeCab の辞書を強化するために引用したデータは, ネットスラングや固有名詞がよく含まれている以下の 3 つである.

1. Wikipedia
2. はてなキーワード
3. ニコニコ大百科データ

Wikipedia はウィキペディア日本語版のダンプ, はてなキーワードではキーワード名とそのふりがなの一覧のデータを使用. ニコニコ大百科データに関しては, 国立情報学研究所が研究者に提供しているもので, ニコニコ大百科に 2014 年 2 月上旬までに投稿された記事すべての記事ヘッダ, 記事本文データとそれに付随する掲示板全データがセットになっている. ニコニコ動画にサービス開始当初から 2016 年 8 月 31 日までに投稿された約 1400 万件の動画のメタデータと, それに対する約 35 億件のコメントデータがセットになっているニコニコ動画コメント等データもあるが, サイズが非常に大きいため, 今回はニコニコ大百科データのみを使用する.[2]

3.2　追加要素

今回着目した単語は “gkbr” である. この単語には主に 2 種類の意味が含まれている.

1. ゴキブリ
2. ガクブル

「ガクブル」とはネットスラングであり, ガクガクブルブルの略で, 恐怖感を表す擬態語として使われている. 取得したツイートに含まれている”gkbr” がこの 2 つのどちらの意味で使われているのかを分析する.

対象となる “gkbr” のデータは,2011 年のニコニコ大百科データの中に「ゴキブリ」として登録されていた.「ガクブル」という意味での “gkbr” は登録されておらず, カタカナ表記の “ガクブル” として登録されている.

Twitter では,「ガクブル」の意味として “gkbr” が使われているため, 手動で追加しなければならない. ここで,「ゴキブリ」と「ガクブル」の使われ方の違いの判断基準がはっきりとしないという問題点が浮上した. 以下は,“gkbr” の使われ方の例である.

～gkbr のゴキブリとしての使われ方～
昨日部屋に gkbr が出た
gkbr は食べられない
～gkbr のガクブルとしての使われ方～
(ひどい事故のニュースを受けて)「gkbr」
gkbr してんのかな,gkbr してた,gkbrってる
gkbr がとまらない, すでに gkbr だよ
怖い gkbr, 毎日 gkbr

「ゴキブリ」としては, 単独で使う場合が多い.「ガクブル」としては, 単体で使われる場合や,「する」などの動詞と組み合わせて使う場合, 単独で使う場合, 文末にくっつけて使う場合などがある.

“gkbr” を「ガクブル」として使う際を考える. 判断基準がはっきりとしていないが,「する」という動詞と組み合わせて使う場合が多いため, とりあえず “gkbr し” として辞書を新たに追加した. 同様に,“gkbr す” として辞書を手動で追加した.

4　結果

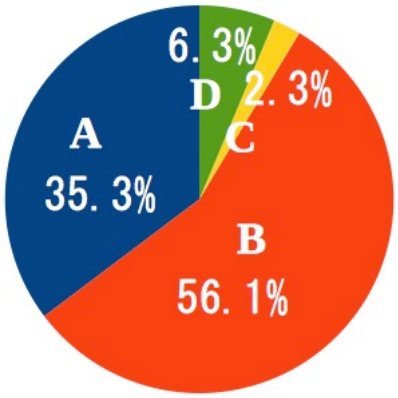
MeCab で “gkbr” を含む文章を解析すると” 未知語” 扱いとされていたものが, 辞書を強化した結果 “gkbr” が名詞として扱われ, そのデータがニコニコ大百科データからとってきたものであるという表示まで可能になった.

ニコニコ大百科データには, 文章がある単語として登録されているものもあるため, そのフレーズを含む文章を解析しようとすると, うまくいかなかった.

文章が単語として登録されている例
でも、そうなんでしょう？
テイルズオブシリーズの声優一覧
なんでこの動画終わらないの？
ガチで痩せそうなダイエット方法　まとめ
聖者は十字架に磔られた

指定した単語 “gkbr” を含む約 2 週間分の 1000 件のツイートをとってきた. その結果, とってきたツイート 1000 件のうち,A は 353 件,B は 561 件,C は 23 件,D は

63 件となった (図 1). ユーザ ID に “gkbr” を入れているユーザは少ないものの, つぶやきの頻度が多いため, 件数が多くなっている. また,“gkbr” は「ゴキブリ」として使われるよりも,「ガクブル」として使われることの方が圧倒的に多い. また,「ゴキブリ」,「ガクブル」どちらの意味で使われているのかが不明なツイートも何件か存在した.



A ユーザ ID に “gkbr” が含まれている
B 「ガクブル」の意味として “gkbr” を含む
C 「ゴキブリ」の意味として “gkbr” を含む
D ユーザ ID, ツイートの両方に “gkbr” を含む または「ガクブル」「ゴキブリ」どちらの意味で使われているか不明

図 1: 取得した 1000 件のツイートにおける割合

「ゴキブリ」か「ガクブル」か意味が判定できない例
味方が gkbr
今時 gkbr なんて誰も使ってないぞ !!!
gkbrってなんて読む ?? wwwwww
gkbr とか懐かしい←
gkbr 発言

「ガクブル」として使う場合において, 手動で追加した “gkbr し” に関しては,「gkbr しね！」という,「ゴキブリ」の意味として使っているツイートをとってきてしまった. そのため,“gkbr し” という辞書は使えないという結果となった.

また,“gkbr が” としても, 直後に「とまらない」とすれば「ガクブル」の意味になり,「出た」とすれば「ゴキブリ」の意味にとれる.

5 今後の課題

ニコニコ大百科データが 2014 年 2 月上旬までに投稿された記事データのため, それ以降に投稿された記事データの反映ができていない.3 年の間に新しく投稿されたデータ, 新しく追加された単語の解析もできるように対応させたい.

ある単語を含むツイートをとってくる際に, 本文に指定の単語が含まれているツイート以外にも, ユーザ ID に指定の単語が含まれているユーザのツイートをすべてとってきてしまうため, ツイート本文にのみ指定の単語を含むもののみをとってこれるようにしたい.

“gkbr” が「ゴキブリ」,「ガクブル」どちらの意味として使われているのかの判断基準を明らかにできていないため, 今回手動で辞書を追加した “gkbr す” 以外にも細かく分類し, 改善したい.

参考文献

[1] https://www.python.jp/

[2] https://ja.wikipedia.org/wiki/Wikipedia:%E3%83%87%E3%83%BC%E3%82%BF%E3%83%99%E3%83%BC%E3%82%B9%E3%83%80%E3%82%A6%E3%83%B3%E3%83%AD%E3%83%BC%E3%83%89http://developer.hatena.ne.jp/ja/documents/keyword/misc/catalog
http://www.nii.ac.jp/dsc/idr/nico/nico.html
http://qiita.com/zaru/items/19f2ba007b46fbc587ed