IML#5 叫喚フレーズによる感動の大きいツイートの抽出

2016/08/25
sunouchi

やりたいこと

- ツイートを感情分析したい
- 感動の大きいツイートを抽出したい

前回の試み：感情語辞書をつかった感情分析

- 感情語辞書の作成（例：['良い, 0.6]）
- 感情抽出処理
    - 語レベル解析(e.g. 良い)
    - 句レベル解析(e.g. とても良い)
    - 文レベル解析(e.g. 梅雨が開けたので、今日は天気がとても良い)

失敗...(⊃Д｀ｏ)ﾉ

- 前回試みた手法は、「文法に則った構造を持つテキスト」には有効である。新聞記事や小説など
- しかし、ツイッターなどのSNSテキストは「文法に則った構造を持たないテキスト」のため有効でない。形態素解析の誤判定、係り受け解析の誤判定などが起きる
- 旧来的な自然言語処理の限界

注目する失敗例

- 「年賀状印刷ミスったぁぁぁ！！」（崩れた表記）
- 「分からないことはググれ」（未知語）

崩れた表記と未知語

- 崩れた表記
    - 例：「年賀状印刷ミスったぁぁぁ！！」
    - 「ミスをした」ではなく「ミスったぁぁぁ！！」になっていて、正しく形態素解析できない
- 未知語
    - 例：「分からないことはググれ」
    - 「ググる」という言葉が辞書に載っていない

崩れた表記に着目する

崩れた表記の分類

- 長音記号の挿入
  - 例：でーす、もしもーし
- 母音の挿入
  - 例：やったあ、行けえええ
- 小書き文字の挿入
  - 例：見たぁぁい、ねむぅい（眠い）
- 促音・発音の挿入
  - 例：すっばらしい、すんばらしい
- 長音記号による置換
  - 例：ありがとー、ねーさん（姉さん）
- 小書き文字による置換
  - 例：ぉぃしい（おいしい）、ヵわいい（かわいい）
- 類似記号による置換
  - 例：こωばωゎ（こんばんは）

崩れた表記は感動の大きさを表しやすい

- 「ありがとおおおおおお！」
- 「帰りたいいいいいいっっ！！！」
- 「それはらめえええええええ」

叫喚フレーズとして定義する

叫喚フレーズの定義

- 語尾の母音が3回以上繰り返して付加されている
- 母音は大文字、小文字を区別しない
- 母音はひらがな、カタカナの大小文字すべて

叫喚フレーズを抽出する正規表現

あ{3,}|い{3,}|う{3,}|え{3,}|お{3,}|ぁ{3,}|ぃ{3,}|ぅ{3,}|ぇ{3,}|ぉ{3,}|ア{3,}|イ{3,}|ウ{3,}|エ{3,}|オ{3,}|ァ{3,}|ィ{3,}|ゥ{3,}|ェ{3,}|ォ{3,}

叫喚フレーズを含んだツイートの分類

- 外因による叫喚（外因）
  - TV番組などの他メディアに対する叫喚や、実世界のイベント・事象に対する叫喚
  - 「あすかあああああああああ #エヴァ」
  - 「AKBくるうううううう〜！」
- 内因による叫喚（内因）
  - 自身の生理的な現象に対する叫喚や、奮起によって生じた叫喚
  - 「うぁーー眠いいいいいいいてか、雨降っとる！」

    - 「ぬぅおおおおおおおおおおがんばらなくちゃあああああ」
- 会話により生じた叫喚（会話）
    - マイクロブログ上でのユーザ同士の会話によって発生した叫喚
    - 「@wahsing_7 わかるううう！」
    - 「@aniota44 あああ(´・ω・`) ちょっと凹みま すね(´；ω；`)」
- 判断不可（不明）
    - メッセージ単体からは判断できない叫喚
    - 「ふおおおおお」
    - 「ひぃいぃぃぃいいぃいいいぃぃいぃぃぃぃ」
- その他・非叫喚（その他）
    - 分類1-4に分類されないツイート、または叫喚が確認できないツイート
    - 「メリークルスニク！！ #メリークルスニクということでクルスニク一族クラスタさん集まれえええ」
    - 「┗(^o^)┛ イタ電するぞぉぉぉｗｗ(^o^)?┓ もしもしｗｗｗｗｗｗｗｗ」

頻出するのは「会話」と「外因」

- 会話　：38%
- 外因　：36%
- 内因　：11%
- 不明　：9%
- その他：6%

叫喚フレーズを含んだツイートの抽出手順

- 1, 前処理としてメンション（@username）、ハッシュタグ、URL、日本語以外の文字、記号をメッセージから除去する
- 2, 叫喚フレーズを含んだツイートを、先に定義した正規表現を用いて抽出する
    - 「年賀状印刷ミスったぁぁぁ」
    - 「うわぁぁぁ最悪だめだぁぁぁ」
- 3, 繰り返し母音を大文字化
    - 「年賀状印刷ミスったあああ」
    - 「うわあああ最悪だめだあああ」
- 4, すべての繰り返し母音部分に対して、母音一文字とそれ以前の文字列を抽出
    - 「年賀状印刷ミスったあ」
    - 「うわあああ最悪だめだあ」

参考文献

- 若井 祐樹, 熊本 忠彦, 灘本 明代, "映画に対する実況ツイートの感情抽出手法の提案", 研究報告データベースシステム（DBS） 2013-DBS-158(16), 1-6, 2013-11-19
- 浅井 洋樹, 秋岡 明香, 山名 早人, "きたあああああああああああああああ！！！！！１１：マイクロブログを用いた教師なし叫喚フレーズ抽出", DEIM Forum 2013
- 奥村 学, "マイクロブログマイニングの現在", 信学技報 vol.111, No.427 NLC2011-59, pp.19-24, 2012
- 笹野 遼平, 鍛治 伸裕, "不自然言語処理～枠に収まらない「リアルな」言語処理～ 2.新しい語・崩れた表記の処理", 情報処理 53(3), 211-216, 2012-02-15

- 高橋 雄太, 片岡 義雅, 浅井 洋樹, 山本 祐輔, 秋岡 明香, 山名 早人, "繰り返し表現を含んだ感情的なツイートの抽出", DEIM Forum 2012 C3-6