P2-19

オンラインコミュニティにおける単語頻度の通時的変化を利用した新語リストの獲得

阿部香央莉[1,2]，松田耕史[2]，吉川将司[1,2]，乾健太郎[1,2]（1. 東北大学，2. 理化学研究所）

概要

- 既存システムは日々生み出される新語に対処不可
  例：「あつ森」「エモい」など
- 新語分析基盤を整えるべく、対処したい新語リストの取得を2種類の手法で試みた

データセット

[1] https://www.nii.ac.jp/dsc/idr/nico/

- 年別（2013～2018）に分割されているテキストコーパスを使用して分析

1. ニコニコデータセット[1]（(株)ドワンゴ提供）
2. Twitterデータ（研究室内でクローリング）

手法1: 頻度0を基準とした獲得

- ある年を境に、頻度が0→100より大 になる単語を取得

2013→2014	STAP細胞, SHIROBAKO, 危険ドラッグ, …
2014→2015	安保法案, ねこあつめ, 刀剣男士, デレステ, …
2015→2016	安倍マリオ, ニンテンドースイッチ, …
2016→2017	フレンズなんだね, PUBG, ハンドスピナー, …
2017→2018	西日本豪雨, ボブネミミッミ, …

★ 両方のデータで頻度0から生起した語は、新語として馴染みのあるものが多くみられた

手法2: 時系列クラスタリングによる獲得

- 頻度の時系列データから、徐々に勢力拡大した単語を取得

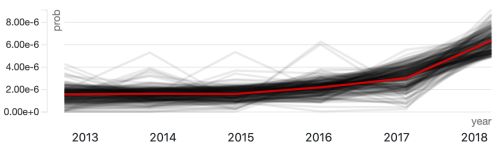

図1: Twitterデータ, 右肩上がり

獲得された単語

認知症, DV, 乱舞, Bluetooth, エモい, アメフト, 皆既月食, AbemaTV, LGBT, プレモル, 田中圭, hour, 暴落, 義援金, 銀メダル, 改憲, 史上最高, …

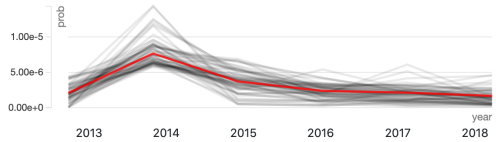

図2: Twitterデータ, 凸型

獲得された単語

アナ雪, 東京喰種, 松岡修造, E-girls, るろ剣, 鬼灯, ギリシャ, 金田一, 寄生獣, ネイマール, 集団的自衛権, ソチ, アルゼンチン, アオハライド, …

クラスタリング結果の詳細はこちら↓
https://chanabe-k.github.io/time_clustering_novel_words/

まとめ・今後の課題

- 対象としたい新語を含む単語集合を獲得できた
  - が、新語リストとして運用するには **人手チェック**が必要
- 新語等の未知語に対処可能な方法論の考案
  - 日本語で未知語を扱う際の問題？
    → **単語分割, 扱う文字の多様性**…