ソフトクラスタリングを導入した
ソーシャルメディア上のトピック抽出

長岡技術科学大学　画像・メディア工学研究室

学部4年　渡部　一翔

SNSでの情報環境の現状

- SNSは日常的に膨大な情報が投稿、共有されており、個人や組織が多彩な立場から情報を発信している。
- 社会に大きな影響を与える情報空間となっている。

このような情報環境においては、膨大な情報を効率的に把握するために意味的なまとまりを持ったトピックに細分化したい.

課題

- 目的の情報を的確に抽出するのは難しい。
- 同じ単語であっても、文脈や時期によって意味合いが変わる場合がある。

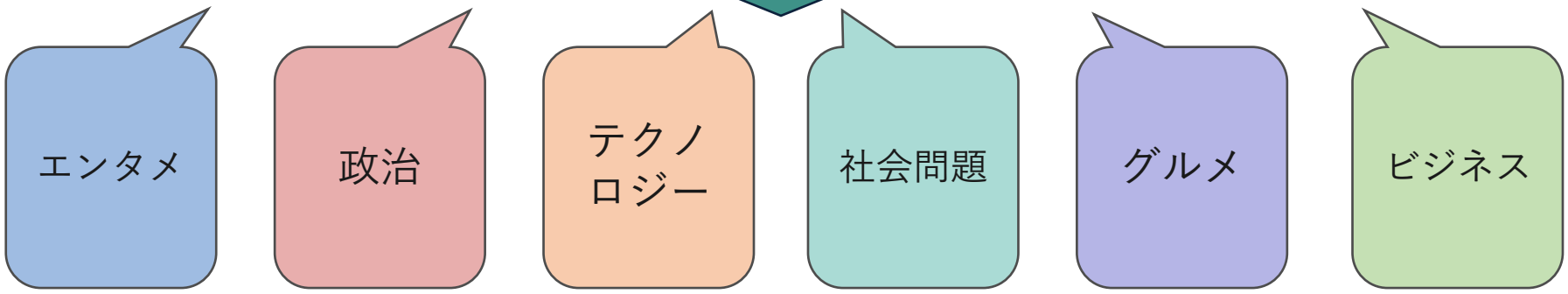
目的

- よって本研究では、SNS上の情報を時系列と意味の両面から解析し、単語の多義性や文脈による変化を考慮したトピッククラスタリング手法の開発を目的とする。

SNSに投稿された単語の集合

（例）お笑い，内閣，半導体，移民，
タピオカ，猛暑，災害，豪雨，など…

クラスタリング



自然言語処理に基づく意味分析手法

BERT[2]

文章内の単語をその文脈から埋め込みベクトルを出力する自然言語処理モデル。

文章を双方向の文脈を同時に考慮することによって単語の意味を高精度にとらえることが可能。

DPGMM [3]

データを複数のクラスタに属させることができるソフトクラスタリングモデル。

従来のGMMを拡張しており、クラスタ数を自動的に推定できるモデル。

etc.

提案手法

GLIPCAとx-meansを用いて、トレンドクラスタをさらに意味ベクトルによりハードクラスタリングを行うことで話題に細分化。[1]

時系列信号処理に基づくトピック検出手法

GLIPCA [4]

間接相関を除去して、トレンドクラスタを生成する時系列クラスタリング手法。

トレンドの時期に注目したクラスタリングが可能。

TV-GLIPCA [5]

従来のGLIPCAにさらに各年の要素を追加したもの。

毎年出るトレンドクラスタ(頻出クラスタ)とたまにしか出ないトレンドクラスタ(希少クラスタ)に分類可能

etc.

[1] Y. Fukai, R. Harakawa, and M. Iwahashi, “ソーシャルメディア上のトピック 抽出のための glipca の高精度化に関する検討,” in 電子情報通信学会信越支 部大会, 2024, pp. 55–55
[2]] J. Devlin et al., “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. Conf. NAACL-HLT, 2019, pp. 4171–4186
[3] D. M. Blei and M. I. Jordan, “Variational inference for dirichlet process mixtures,” Bayesian Analysis, vol. 1, pp. 121–143, 2006
[4] R. Harakawa and M. Iwahashi, “Ranking of importance measures of tweet communities: Application to keyword extraction from COVID-19 tweets in japan,” IEEE Trans. Computational Social Systems, vol. 8, no. 4, pp. 1030–1041, 2021.
[5]R.Mizuma, R.harakawa, and M.Iwahashi,” トレンドクラスタの判別のためのTime-Varying GILPCA-猛暑・残暑に関するツイートの分析,”in 信号処理シンポジウム, 2023, pp. 1-6

時系列クラスタリング

GLIPCA

自然言語トピック検出

BERT

DPGMM

- 時系列だけでは同時期に盛り上がった単語の集まりであるため、その中で類似したトピックのみを抽出することができない。
- 意味ベクトルを使用してクラスタリングすることによって時系列クラスタをトピックに細分化することが可能。
- さらに、ソフトクラスタリングを用いることで単語の多義性を反映したクラスタリングが可能となる。

user1
熱中症に注意しよう！

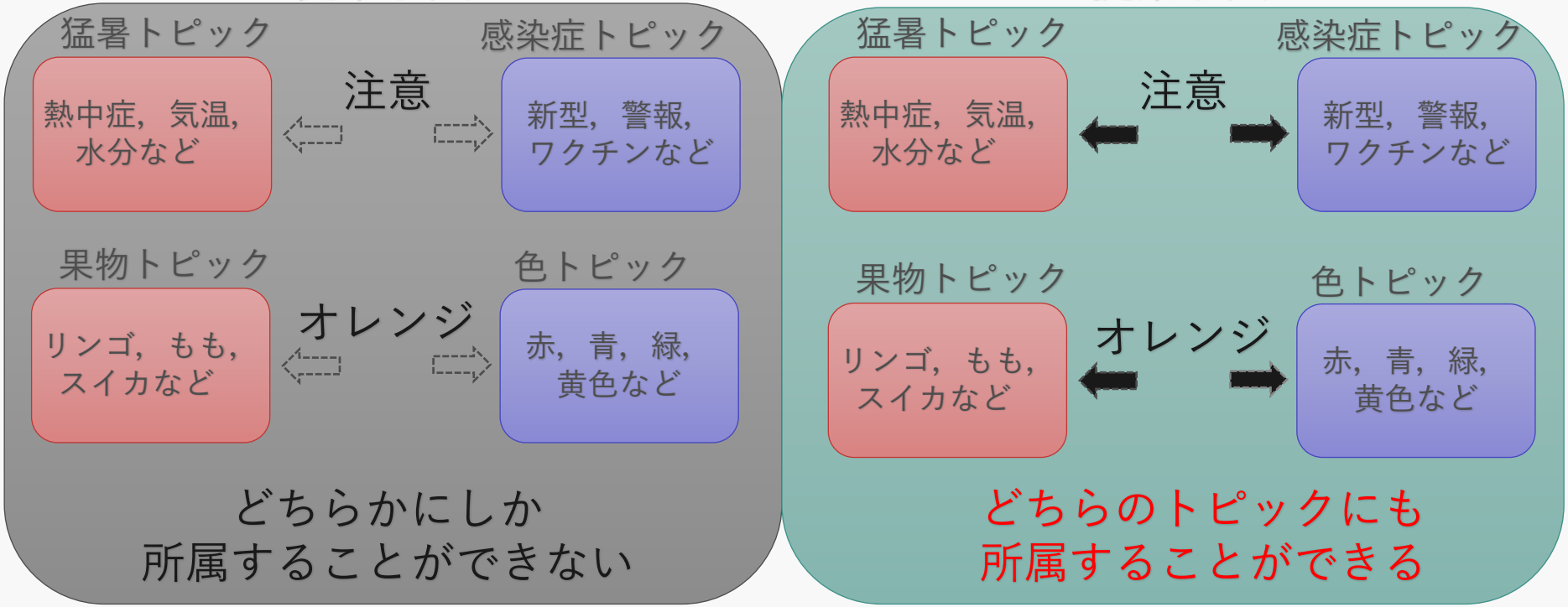
user2
感染症に注意しよう！

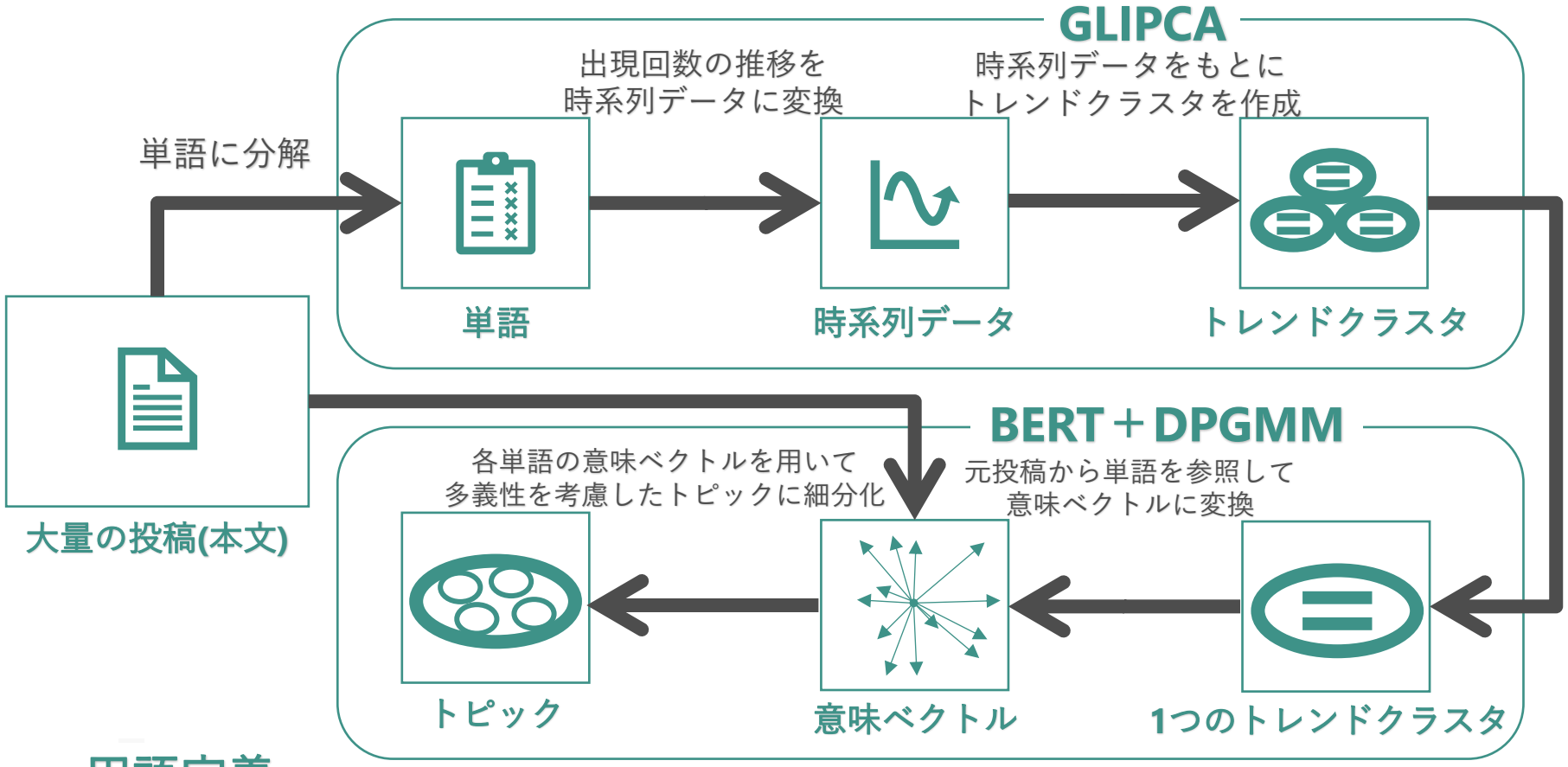
「注意」という単語は様々な話題に出現する

user1
オレンジを使ったデザート

user2
オレンジ色の服

「オレンジ」という単語は様々な話題に出現する





用語定義

- **トレンド** 単語の出現頻度を時系列信号として表現したもの
- **トレンドクラスタ** GLIPCAの出力結果で、トレンドのピークが類似する単語の集合
- **トピック** トレンドのピークが近く、意味も類似する単語の集合

これらの定義は先行研究である深井らの研究[1]で用いられた定義をそのまま採用する。

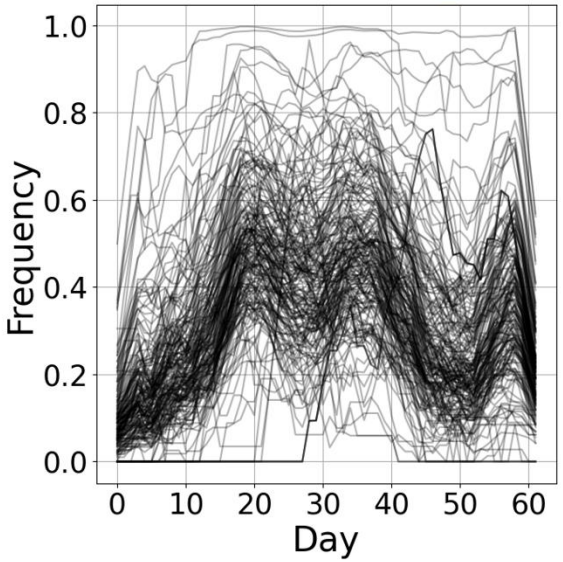
意味ベクトル（埋め込み）

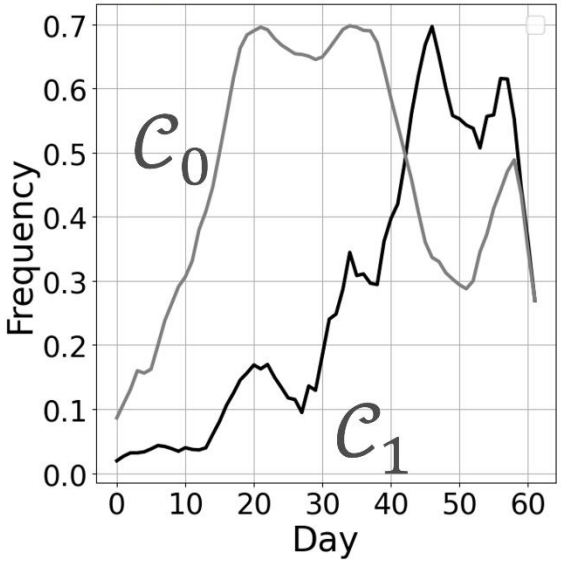
- 自然言語処理において、単語や文章を多次元の数値ベクトルに変換する表現方法
- 単語同士の意味的な類似性をベクトル空間上の距離として計算可能にする。

コサイン類似度

- 複数のベクトルがどれだけ同じ方向を向いているの指標
- 今回はクラスタ内の単語の意味ベクトルがどれだけ類似しているかを測定するために使用する。
- 1に近いほど意味的に近く、0に近いほど無関係となる。

GLIPCA





トレンドクラスタc_0

アスリート，拡大，令和，アメダス，代償，組織，世界，接近，悪化，上陸，読売新聞，政治，デルタ，Japan，観測，宣言，列島，死者，委員，直撃，北海道新聞，地震，ウェザーニュース，新型，etc

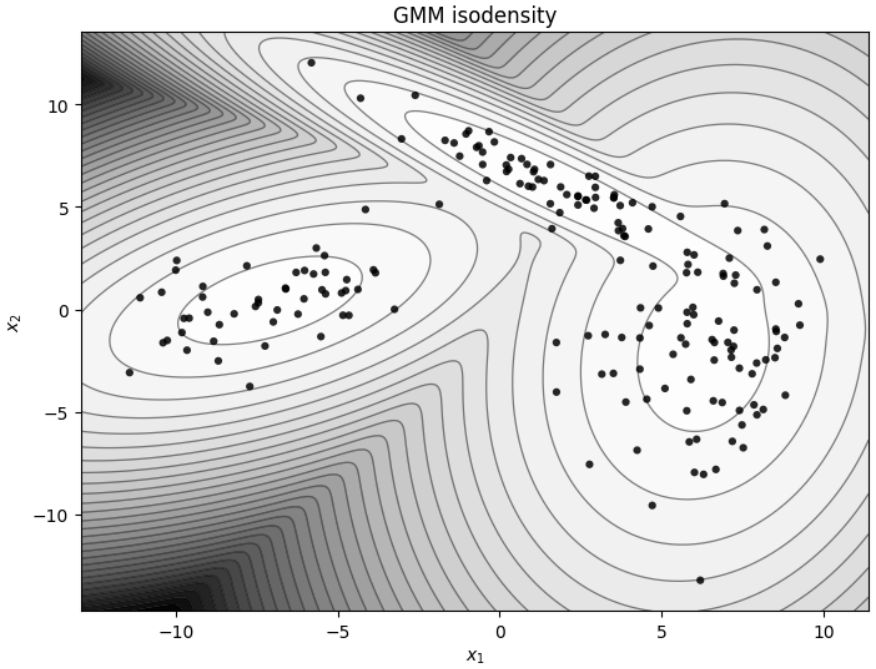
- GLIPCAによってトレンドクラスタを生成可能
- トレンドクラスタとは同時期に盛り上がった単語の集合
- 間接相関を抑制してトレンドクラスタを抽出することが可能

DPGMM

GMMをクラスタ数が自動決定できるように拡張したもの

GMMとは複数の正規分布の組み合わせによってデータ全体の確率分布を表現することが可能

- 単語が各トピックへの所属確率を算出することができる
- ソフトクラスタリング可能
- クラスタ数を自動決定することができる



入力データ

2021年 7月1日から8月31までの

Twitter(現:X)の「猛暑」を含む日本語投稿本文

実験手順と評価法

- GLIPCAを使用してトレンドクラスタを生成する。
- そのうち1番単語数の多いトレンドクラスタ内の単語と元投稿データをもとにBERTにより意味ベクトルを取得する。
- 取得した意味ベクトルのにDPGMMを使用してソフトクラスタリングを行い、トピックに細分化する。
- 得られたトピックのコサイン類似度を算出し、細分化前のトレンドクラスタと比較する。

トピック数は65抽出され、トピックへの所属確率が0.3以上のものを所属としている．複数のトピックに分かれたものは(Multi)とつけている。
下に示すトピックは一部抜粋である

トレンドクラスタ

気象災害トピック
- 干ばつ(Multi)
- 悪化
- 地点(Multi)
- ウェザーニュース
- 死者(Multi)

感染症関連トピック
- 宣言
- 拡大
- 新型
- 令和(Multi)
- デルタ(Multi)

日本五輪トピック
- Japan(Multi)
- アスリート
- 委員
- 代償(Multi)
- 組織

気象報道トピック
- アメダス
- 観測(Multi)
- 地点(Multi)
- 北海道新聞(Multi)
- 読売新聞

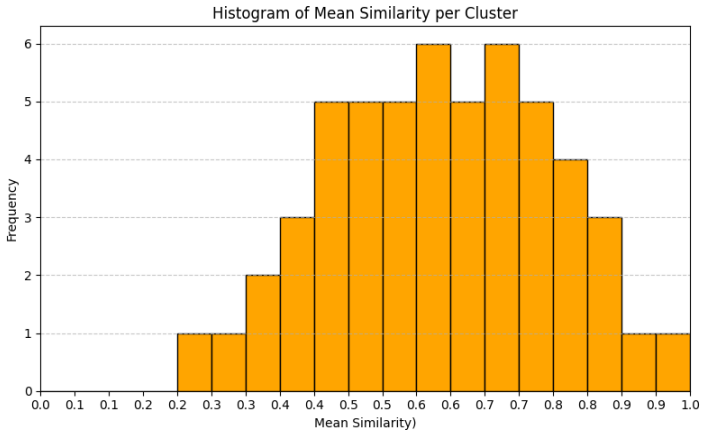
自然災害トピック
- 地震
- 接近
- 列島(Multi)
- 上陸
- 直撃

国際政治トピック
- 世界
- 政権
- 死者(Multi)
- 政治
- 代償(Multi)

- 各トピック内の単語の総当たりコサイン類似度で評価

※この評価指標は[1]のものと同じ方法を使用する

$$D_S = \frac{1}{{}_N C_2} \sum_{1 \le i \le j \le N} \frac{e_{W_i} \cdot e_{W_j}}{\|e_{W_i}\| \left\|e_{W_j}\right\|}$$



W :対象となる単語の集合
N :1つの集合に含まれる単語の数
${}_N C_2$:集合の総当たりの回数
e_{W_i}, e_{W_j} :2つの単語の埋め込みベクトル
D_S :意味的密度
D_Sが高いほどトピックが意味的に類似している

- 平均類似度 0.6130
- 最大類似度 0.9519
- 最小類似度 0.2242

(参考)トレンドクラスタ
平均類似度 0.0165
と類似度が大きく上昇した。

まとめ

- 今回の実験では、SNSの投稿に対して、時系列クラスタリング及び文脈を考慮して意味クラスタリングを使用した多義性を尊重するトピッククラスタリング手法を提案した。
- この手法によって、時系列的ピーク、意味ベクトルの両方が類似したトピックを抽出することができ、同じ単語でも分けて抽出することができた。

今後の展望

- 投稿に付属した画像や動画を現在は考慮できていないため、こちらも考慮したトピック分類を可能にしたい。
- 多義性についての評価が現在できていないため定量的な評価をできるようにしたい。

user1

熱中症に注意しよう！

user2

感染症に注意しよう！

「注意」という単語は様々な話題に出現する．また，

user1

オレンジ色の服を買った

user2

オレンジを使ったスイーツ

「オレンジ」という単語は使われる文脈によって意味が異なる