多義性に対応した GLIPCA トレンドクラスタの文脈細分化

Topic classification by time series signal processing considering polysemy

渡部一翔　深井陽平　原川良介　岩橋政宏

Ichito Watanabe　Yohei Fukai　Ryosuke Harakawa　Masahiro Iwahashi

長岡技術科学大学

Nagaoka University of Technology

1．　まえがき

近年、SNS は災害や社会課題に対する人々の関心をリアルタイムには反映する情報源である。しかし、そこに現れる単語は、一見同じでも、時期や文脈によって、意味が異なる場合がある。例えば「落ちる」という単語は、ある文脈では物理的な落下を、ある文脈では不合格を指す。本研究では、このような単語の多義性や意味の変化をとらえ、時系列(単語の日ごとの登場頻度の変化)及び文脈を考慮した精密な話題抽出を行うことを目的とする。

本文では、日ごとの投稿頻度が類似する単語の集合をトレンドクラスタ、その中でも意味が類似する単語の集合を話題と定義する。

トレンドクラスタを抽出する従来方法として Graphical Lasso-Guided Principal Component Analysis (GLIPCA)[1]がある。GLIPCA では、スパーズ構造学習に基づき間接相関を抑制することで、トレンドクラスタ$\mathcal{C}^n(n=1,2,\dots\mathcal{C};$ $\mathcal{C}$は得られた単語の集合の数)を抽出している。

本研究では、トレンドクラスタ$\mathcal{C}^n$に対して、文脈を考慮したソフトクラスタリングを行う手法を提案する。先行研究として、深井ら[2]の時系列及び自然言語処理を用いた話題抽出があるが、本研究では、文脈を考慮したソフトクラスタリングを行うことで、先行研究では困難であった単語の多義性を考慮した話題抽出が可能となる。

2．提案手法

入力するデータは 2021 年 7 月 1 日から 8 月 31 日までの「猛暑」を含む Twitter(現 X)の投稿本文(※広告等を含むデータは除外)である。入力データDに対して、GLIPCA を適応し、トレンドクラスタ$\mathcal{C}^n$を抽出する。$\mathcal{C}^n$内の各単語について、もと投稿データを参照し、BERT を使って、文脈を考慮した埋め込み$e_w(w \in \mathcal{C}^n)$を生成し、ソフトクラスタリングを行う。今回使用するクラスタリング手法は Dirichlet Process Gaussian Mixture Model(DPGMM)[3]であり、この手法は正規分布の混合によってデータをクラスタリングする GMM[4]を拡張し、クラスタ数を事前に決めることない非パラメトリックな手法である。DPGMM を単語の集合に使用すると、各単語のクラスタ所属確率$P(z_k|e_w)$が取得できる。なおここでのz_kは DPGMM により得られた k 番目の話題を指す。同一の単語 w に対して、複数の文脈に基づく異なる埋め込みベクトル$e_w^{(m)}(m=1,2,\dots,M;M_w$は単語$w$に対する埋め込みベクトルの個数) を得る。 $e_w^{(m)}$より$P(z_k|e_w^{(m)})$を算出し、単語 w における総合的なクラスタ所属確率$\bar{P}(z_k|w)=\frac{1}{M_w}\sum_{m=1}^{M_w}P(z_k|e_w^{(m)})$を定義する。これにより、同一単語が異なる文脈で異なる意味を持つ場合においても、その多義的な出現状況を定量的に捉えることが可能となる。

3．　実験

提案手法によって抽出された話題が、本研究の目的するものとなっているかを検証するため、先行研究である[2]の意味的密度$D_s=\frac{1}{{}_NC_2}\sum_{1\le i<j\le N}\frac{e_{w_i}\cdot e_{w_j}}{\|e_{w_i}\|\|e_{w_j}\|}$を算出する。

ここでは、N は単語の集合$\mathcal{W}$に含まれる単語数である。同一単語の異なる埋め込み$e_w^{(m)}$がある場合は、単語の集合の重心に一番近いものを利用する。

意味的密度D_sが高いほど単語の集合$\mathcal{W}$内の単語が意味的に類似していることを示す。

図 1：トレンドクラスタから細分化された話題

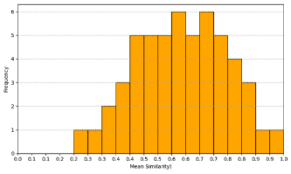

図 2：意味的密度D_sの分布

図 1 に提案手法によって。トレンドクラスタを細分化した話題に含まれる単語の一部を示す。話題は全 70 話題抽出された。これを見ると、内容のまとまりがある話題に細分化せれており、さらに同一の単語が複数の話題にまたがって出力されていることがわかる。

図 2 に各話題の意味的密度D_sの分布をヒストグラムで示す。抽出された話題の意味的密度D_sの平均は 0.613 となり、最大値は 0.952となり、細分化前のトレンドクラスタのD_sが 0.016 であった。このことから本研究で目標とした多義性が考慮して、意味的に類似する話題を抽出できたことを示している。

参考文献

[1] Harakawa R, Iwahashi M. Ranking of Importance Measures of Tweet Communities: Application to Keyword Extraction From COVID-19 Tweets in Japan. IEEE Trans Comput Soc Syst. 2021

[2] 深井, 原川, 岩橋. 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. 電子情報通信学会信越支部大会. 2024

[3] C.E. Rasmussen The Infinite Gaussian Mixture ModelIn: Advances in Neural Information Processing Systems (NeurIPS), vol.12.2000

[4] D.M. Blei, M.I. Jordan, Variational inference for Dirichlet process mixtures. Bayesian Anal. 1.2006