

経時的テキストデータにおける 2 値の共変量の効果の可視化

滋賀大学 和泉志津恵、県立広島大学 富田哲治、広島大学 川野徳幸、佐藤健一

1. はじめに

近年、注目されているビッグデータの一つに経時的に観測されたテキストデータがある。本研究では、変化係数モデルを用いて、2 値の共変量が付いた経時的テキストデータにおける共変量の効果を可視化する方法を提案する。

2. 経時的テキストデータにおける 2 値の共変量の効果の可視化

和泉ら（2015）や富田ら（2017）の方法論の拡張を検討する。まず、時間に関して線形な変化係数を伴うポアソン回帰モデルを用いて、テキストデータに含まれるキーワードの時系列的な出現頻度に対する共変量の効果を推測し、頻度を予測する（下図の左を参照）。次に、観測時点の始点、終点、節点における共変量効果の推定値を用いて、共変量効果の経時変化のパターンを分類する（下図の中央を参照）。散布図の横軸に共変量の効果の推定値、縦軸に頻度の予測値の最大値を取り、分類結果に基づいて色付けしたキーワードを、観測時点ごとに可視化し、統計解析ソフト R でアニメーション化する（下図の右はアニメーションのスナップショットである）。

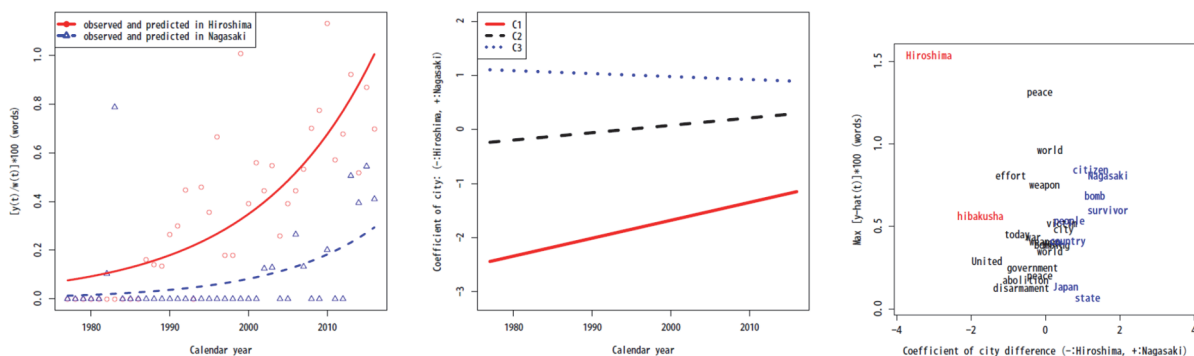


図. 英語版の平和宣言データ[1]のキーワードに対する提案手法による頻度の予測曲線（左）、共変量の効果の分類結果（中央）、共変量の効果の推定値とキーワードの頻度の予測値（右）

[1] Izumi, S., Tonda, T., Kawano, N., Satoh, K. Estimating and visualizing the time-varying effects of a binary covariate on longitudinal big text data. The International Journal of Networked and Distributed Computing, 5(4), 243-253, 2017.

[2] 富田哲治, 佐藤健一, 和泉志津恵. 広島平和宣言における単語出現頻度に基づく広島の平和観の経時変化について. 長崎医学会雑誌, 91(特集号), 176-179, 2017.

[3] 和泉志津恵, 佐藤健一, 川野徳幸. 経時的に観測されたテキストデータに対する 変化係数モデルに基づく統計的な分類方法と視覚化について. 計算機統計学, 28(1), 81-92, 2015.

連絡先: 和泉 志津恵 (Email: shizue-izumi@biwako.shiga-u.ac.jp)