Semantic Shift Stability: 学習コーパス内の単語の意味変化を用いた事前学習済みモデルの時系列性能劣化の監査

石原祥太郎 [1]　高橋寛武 [2]　白井穂乃 [1]
[1] 株式会社日本経済新聞社　[2] 独立研究者
shotaro.ishihara@nex.nikkei.com

掲載号の情報

31 巻 4 号 pp. 1563-1597.
doi: https://doi.org/10.5715/jnlp.31.1563

概要

研究者や実務者にとって事前学習済みモデルの利活用が一般的になる中，独自モデルを構築・運用する際には，事前学習時に存在しなかった新しいテキストに対する性能劣化に注意しなければならない．時系列性能劣化[1)]を計測する最も素朴な方法は，実際に新しいテキストを学習コーパスを加えて事前学習済みモデルを構築し，ファインチューニングをして新しいテキストに対する推論を行い，性能を評価し比較することである（図 1 上）．しかし，大規模な事前学習済みモデルの構築・ファインチューニング・推論には膨大な計算量が必要なため，費用や時間の側面が実用上の課題となる．

本研究の目的は，事前学習済みモデルの時系列性能劣化を，事前学習・ファインチューニング・推論なしに監査する枠組みの開発である（図 1 下）．我々は事前学習済みモデルの時系列性能劣化は学習コーパス内の単語の通時的な意味変化に起因するという仮説を定め，セマンティックシフトの研究領域の知見を応用した監査指標 **Semantic Shift Stability** を設計した．この指標は，異なる時間幅の学習コーパスで作成された 2 つの word2vec モデルを比較することで計算される．word2vec モデルは比較的低コストで作成できるため，この指標から時系列性能劣化を推測できれば，優れた監査の仕組みとなる．

提案する監査の枠組みの有用性を検証するため，予備実験として日本語の RoBERTa モデルと日本語・英語の word2vec モデルを学習コーパスの期間を変えてそれぞれ 11 個ずつ作成し，事前学習済みモデルの時系列性能劣化を観察した．その後，これらのモデルの時系列性能劣化を監査するという設定で，Semantic Shift Stability の有効性を検証した．算出された値の年次推移を図 2 に示す．実験を通じて，学習用コーパス間の単語の通時的な意味変化が大きい 2016 年や 2020 年に，事前学習済みモデルにも大きな時系列性能劣化が発生していると分かった．Semantic Shift Stability の利点を活かし，意味が大きく変化した単語から原因を推察した結果，2016 年の米大統領選や 2020 年の新型コロナウイルス感染症の影響が示唆された．

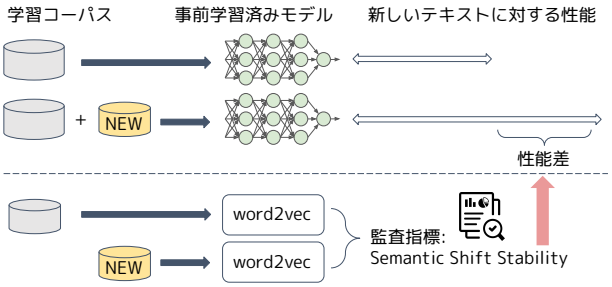

図 1　本研究の問題設定の概要図．学習コーパス内の単語の通時的な意味変化に着目し，事前学習済みモデルの時系列性能劣化を，実際に計測する以前に監査するための手法を開発する．

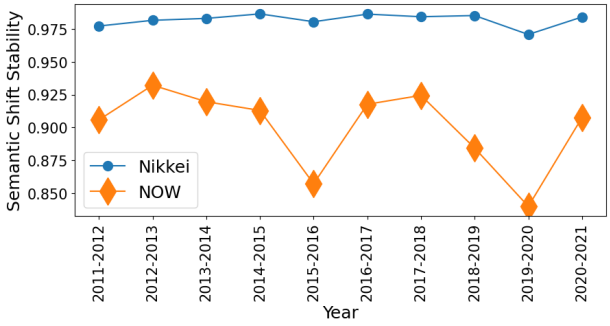

図 2　Semantic Shift Stability の年次推移．日本語（Nikkei）と英語（NOW）の両者で 2019 年と 2020 年の値が最も小さく，相関係数 0.66 と変遷が類似している．

1)　本研究では，事前学習済みモデルの性能が，事前学習時に存在しなかった新しいテキストに対して劣化する現象を時系列性能劣化と呼ぶ．