大域的な類似度と部分文字列を用いた未知語分散表現の生成手法

五十川真生 †　梶原智之 ‡　荒瀬由紀 †
大阪大学大学院情報科学研究科 † ，大阪大学データビリティフロンティア機構 ‡

研究背景

研究背景

自然言語処理タスクにおいて
単語分散表現が有用

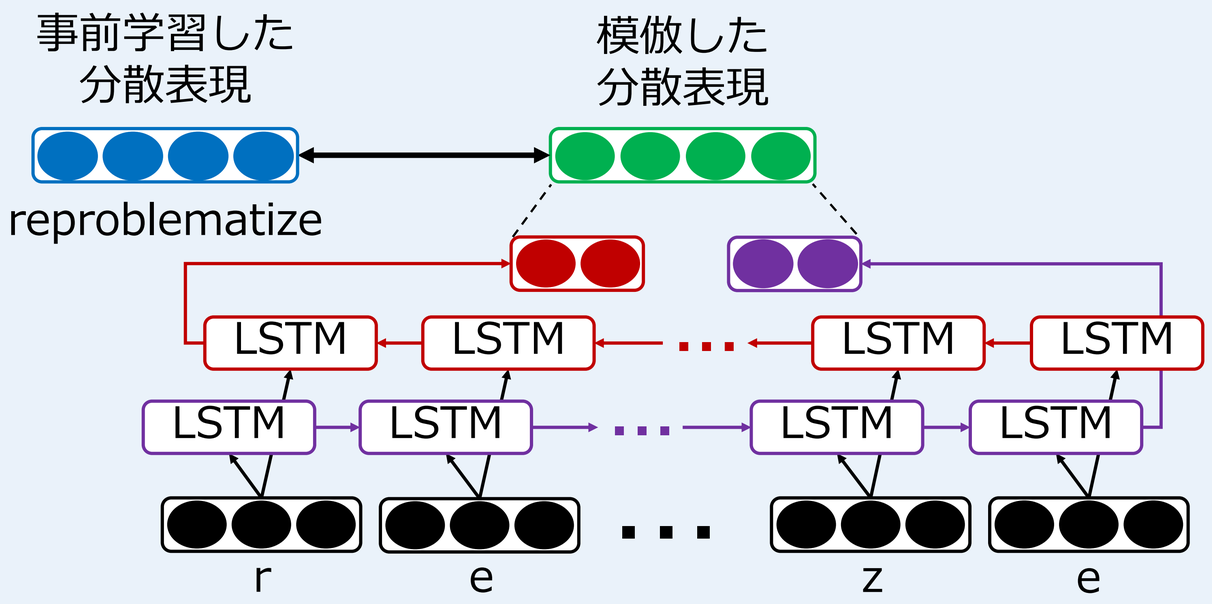
問題点

未知語の分散表現の生成が困難

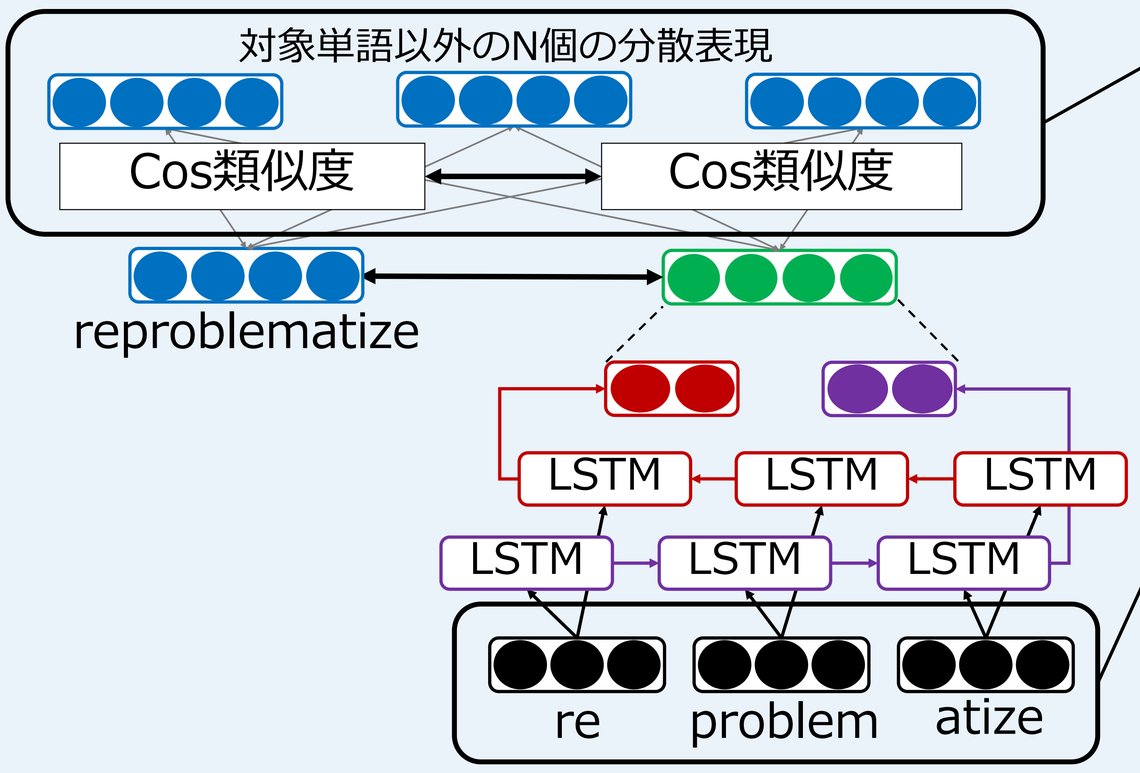
既存研究（Mimick）*1

文字ベクトルから単語ベクトルを生成

*1 Pinter et al, Mimicking Word Embeddings using Subword RNNs, EMNLP, 2017



提案手法



大域的な学習手法

- 単語の意味は他の単語との関係から決定
- 単語そのものではなく**他の単語との関係**を模倣
- N個の分散表現はfastTextから選択
  - 半分は類似度が高い順に選択
  - 残り半分はランダム

部分文字列の利用

単語の意味は**部分文字列**によって構成

reproblematize → re problem atize
再度問題視する　再度　問題　見つける

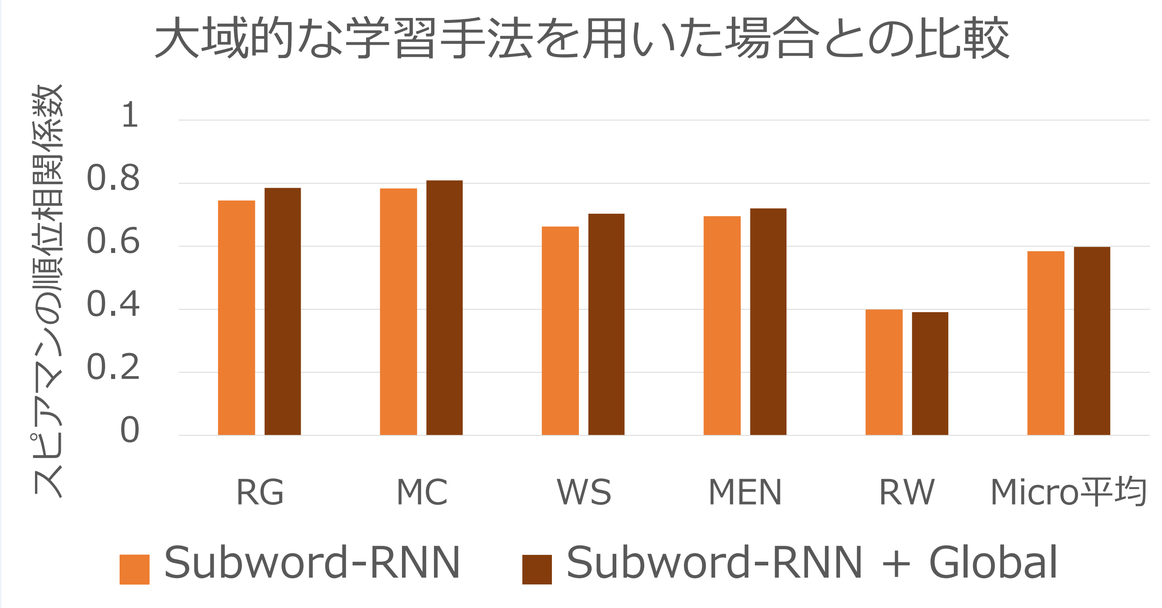
評価実験：単語間類似度推定タスク

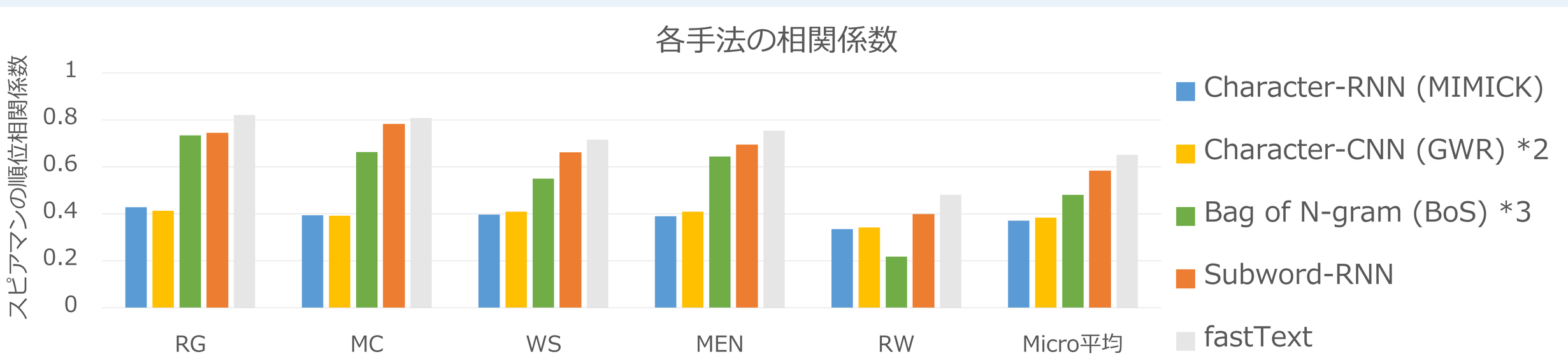
実験設定

- fastTextによって生成した頻度上位10万語の分散表現を学習データ(内1000件は開発データ)に使用
- 提案手法ではN=100に設定

実験結果

- 部分文字列を用いることで性能が向上
- 大域的な学習手法を用いることで性能が向上





*2 Kim et al, Learning to Generate Word Representations using Subword Information, COLING, 2018
*3 Zhao et al, Generalizing Word Embeddings using Bag of Subwords, EMNLP, 2018