



HOKKAIDO UNIVERSITY

Title	Study on Fusion of Distributional Representations and Distributed Representations [an abstract of dissertation and a summary of dissertation review]
Author(s)	劉, 曉冬
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第16176号
Issue Date	2024-09-25
Doc URL	https://hdl.handle.net/2115/93518
Rights(URL)	https://creativecommons.org/licenses/by/4.0/
Type	doctoral thesis
File Information	Liu_Xiaodong_review.pdf, 審査の要旨



学位論文審査の要旨

博士の専攻分野の名称 博士 (情報科学)

氏名 劉 曉冬

審査担当者 主 査 准教授 伊藤 敏彦

副 査 特任教授 坂本 雄児

副 査 教 授 長谷山 美紀

副 査 教 授 土橋 宜典

副 査 教 授 小川 貴弘

学位論文題名

Study on Fusion of Distributional Representations and Distributed Representations

(二種類の分散表現の融合に関する研究)

本論文では、NLP での分布意味論と分散意味論の融合が実世界のアプリケーションや意味表現における NLP の発展に寄与するかどうかを調査することに焦点を当てている。

NLP の分野ではトランスフォーマーベースのアーキテクチャが BERT のようなオートエンコーディングモデルや GPT のような自己回帰モデルといった既存の強力な言語モデルに大きな影響を与えている。これらのニューラルモデルは大規模言語モデル (LLM) として知られ、学術研究の方向性だけでなく日常生活にも影響を与えている。

LLM から出力される意味表現は、Word2Vec や ELMo といった非トランスフォーマーベースのニューラル埋め込みモデルと同じく分散表現である。分散表現は密なベクトル表現であり、すべてのベクトルの各次元が異なる単語やトークンに対して数値で分布される意味軸を表している。これらの意味情報は、各単語やトークンごとにベクトル全体に分散する。

一方、分布表現は疎なカウントベクトルであり、特定の文書や文における各単語の出現頻度が特徴として使用される。分布意味論の実装はベクトル空間モデル (VSM) として知られており、用語文書行列における頻度分布を用いた意味類似度の測定を通じて、情報検索や文書分類によく使用される。分散表現と分布表現には共通点が二つある。第一に、分布表現の用語文書行列が密なベクトル表現に因数分解されると、分布表現はある程度分散表現と同等になる。第二に、分布仮説があり、「単語の意味はその共起する単語によってわかる」という考えである。この仮説は BERT の Masked Language Model (MLM) 事前訓練目標や Word2Vec の Continuous Bag of Words (CBOW) 事前訓練目標に採用されている。一方、分布表現もこの仮説に基づいており、ベクトル空間における分布表現のコサイン類似度によって意味の類似度を測定することができる。

本論文では、二つの融合アプローチ手法「マッピング」と「構造的接続」を提案している。マッピングアプローチでは、著者は重み付けされた分散表現を二つの訓練可能な射影行列を介して分布表現にマッピングする新しい方法を提案している。訓練された行列は推論中に固定され、その結果、異なるタスクデータセットからの任意の文ペアが分散表現からの射影を介して分布意味表現を取得し、取得された表現はパラフレーズ識別やその他の NLP 文ペアタスクのための異なる分類器への入力として使用される。この場合、異なるタスクのために最適化する必要があるのは分類器のパラメータのみとなる。このアプローチによって、文ペアタスクのパワー効率の高い継続学習に大きな寄与す

ることが確認できた。

構造的接続アプローチでは、著者は見出し語検索の考え方から作成された分布表現に DG 埋め込みを提案した。DG 埋め込みは、BERT の MLM 事前訓練目標におけるターゲット単語予測のために双方向の文脈単語を定義する。その結果、事前訓練後の BERT においてより良い語彙習得が可能となった。また、分布意味論の特性により、将来の計算コストが低い拡張の可能性が示唆された。本論文は以下の 5 つの章から構成される。

第 1 章では、分布表現と分散表現とはどのようなものか、それらを融合し利用することの目的、その融合によって解決できる NLP 研究の課題について述べている。

第 2 章では、表現を融合することに関する著者の提案手法に関連した先行研究について紹介している。

第 3 章では、NLP タスクの計算コストが高い継続学習の問題に対処するための提案手法、マッピングによる融合について説明している。実験結果は、タスクにおいて adapter-BERT と近い性能を出しながら、必要なトレーニングのデータ量はわずか 16% 程度、パラメータ更新のための時間は 69-96% 節約可能であることを述べている。

第 4 章では、言語モデルの異方性意味表現の問題に対処するための提案手法、構造的接続による融合について述べている。実験結果は、著者の手法に基づく事前学習後に単語埋め込みのコーンの開口角が拡大することを示した。同時に、事前学習モデルは NLP のタスクにおいて大幅な改善を達成したことも示している。

第 5 章では、本論文のまとめに加えて、研究に関する将来の展望についても述べている。

これを要するに、分布表現と分散表現は相補的な関係にあり、前者は知識エンティティ、後者は一般的な単語に対して有利であることから、著者はこれらをマッピングや構造的な接続を通じて融合させる試みを行った。その結果、NLP タスクにおいて資源コスト効率の高い言語モデルと、意味的に堅牢な言語モデルを実現させた。これにより、計算資源とメモリ資源を低く抑えつつ良好なタスクパフォーマンスを維持したり、意味的に堅牢な言語モデルにより優れたタスクパフォーマンスを発揮することを可能にしたことは、自然言語処理工学の発展に貢献するところ大なるものがある。よって著者は、北海道大学博士 (情報科学) の学位を授与される資格あるものと認める。