数理・情報のフロンティア
2020 年度採択研究者

2021 年度
年次報告書

横井 祥

東北大学 大学院情報科学研究科
助教

言葉が埋め込まれた空間の形と言葉の意味の接続

§1. 研究成果の概要

単語埋め込み空間の幾何学的な特徴および文の表現方法に関する研究を行った。特に 3 つの研究に関して概要を述べる。

[Kobayashi, Kuribayashi, Yokoi, Inui; EMNLP 2022] トランスフォーマーと呼ばれる深層学習アーキテクチャが単語の埋め込み表現を更新する際に文脈情報がいかに混ぜ合わせられているかを定量的・定性的に評価する方法を提供した。注意機構で混ぜ合わせられた文脈情報が、残差結合および層正規化でそのほとんどがキャンセルされること、また BERT と呼ばれる穴埋め言語モデルは高頻度語を無視する(周辺単語の情報によって上書きされやすい)ことがわかった。

[石橋, 横井, 須藤, 中村; NLP 2022] 文をはじめとした単語集合に対する集合演算を単語の意味的類似性を反映した単語埋め込み空間上で実現するためのフレームワークを提案した。提案法では単語集合を単語埋込空間の線形部分空間で表現し、量子論理の枠組みに基づき集合間演算を実現する。提案したフレームワークによって近い概念を示す単語集合を拡張するタスクや文間の意味類似度を推定するタスクで高い性能が達成された。

[大山, 横井, 下平; NLP 2021] 単語ベクトルの長さが単語の意味の強さを表すことを理論的・経験的に示した。はじめに対照学習を用いた典型的な学習アルゴリズムから得られる単語ベクトル空間を情報幾何の言葉で特徴づけ、単語ベクトルのノルムと「その単語が存在することで周辺単語の分布がどの程度歪むか」が対応することを示した。また、単語頻度の交絡を除去してもこの性質が担保されていることを経験的に示した。

【代表的な原著論文情報】

1) Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, Kentaro Inui. “Incorporating Residual and Normalization Layers into Analysis of Masked Language Models.” In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.4547-4568, 2021.
2) Hiroki Ouchi, Jun Suzuki, Sosuke Kobayashi, Sho Yokoi, Tatsuki Kuribayashi, Masashi Yoshikawa, Kentaro Inui. “Instance-Based Neural Dependency Parsing.” Transactions of the Association for Computational Linguistics (TACL), Vol. 9, pp.1493-1507, 2021.
3) 小林 颯介, 横井 祥, 鈴木 潤, 乾 健太郎. “訓練事例の影響の軽量な推定.” 自然言語処理, Vol. 28, No. 2, pp.573-597. 2021.
4) Ayato Toyokuni, Sho Yokoi, Hisashi Kashima, Makoto Yamada. “Computationally Efficient Wasserstein Loss for Structured Labels.” In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pp.1-7, 2021.

5) 石橋 陽一, 横井 祥, 須藤 克仁, 中村 哲. “線型部分空間に基づく学習済み単語埋込空間上の集合演算.” 言語処理学会第 28 回年次大会 (NLP), Online, March 2022.
6) *大山 百々勢, *横井 祥, *下平 英寿. “単語ベクトルの長さは意味の強さを表す.” 言語処理学会第 28 回年次大会 (NLP), Online, March 2022. (* equally contributed)