様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月　 ８ 日現在

機関番号：１４３０１
研究種目：基盤研究(B)（一般）
研究期間：2020～2022
課題番号：２０Ｈ０４１４８

研究課題名（和文）多ドメイン関連性データのグラフ埋め込みによる表現学習

研究課題名（英文）Representation learning through graph embedding of multi-domain relational data

研究代表者
下平　英寿（SHIMODAIRA, Hidetoshi）

京都大学・情報学研究科・教授

研究者番号：００２９０８６７

交付決定額（研究期間全体）：（直接経費）　　13,700,000円

研究成果の概要（和文）：関連性データの埋め込みがどのように情報を表現しているかを理解するための研究を行った．具体的には，埋め込んだベクトルの加減算によるアナロジー計算の基礎となる加法構成性や，それに関係した埋め込みの性質を調べた．通常の加法構成性ではベクトルの和によって意味が同時に成立すること（AND）を表すが，意味のどちらかが成立すること（OR）は頻度重み付き重心，意味の否定（NOT）は対象単語集合の重心に原点を取り直した上で負の方向であることを示した．また，一種の対照学習（SGNS）で得られる単語ベクトルのノルムの２乗がカルバック・ライブラー（KL）情報量で近似され「意味の強さ」を表すことを理論と実験で示した．

研究成果の学術的意義や社会的意義
本研究は，関連性データの埋め込みと表現学習に関する新たな知見を提供しました．加法構成性や埋め込みの性質に関する結果は，単語や概念をベクトルで表現する方法に関する理論的な理解を深めることに貢献しました．これにより，なぜニューラルネットが効果的に機能するのか，その原理を理解する道を開くことが期待されます．

研究成果の概要（英文）：We conducted research to understand how embeddings of relational data represent information. Specifically, We investigated the additive compositionality that forms the basis of analogy calculations using embedded vectors and examined the properties of the embeddings related to it. In conventional additive compositionality, the sum of vectors represents the simultaneous existence of both meanings (AND). However, I demonstrated that the existence of either meaning (OR) can be represented by a frequency-weighted centroid, and the negation of meaning (NOT) is indicated by the negative direction when the origin is relocated to the centroid of the target word set. Furthermore, We theoretically and experimentally demonstrated that the squared norm of word vectors obtained through a form of contrastive learning (SGNS) can be approximated by the Kullback-Leibler (KL) divergence and represents the "strength of meaning."

研究分野：統計学と機械学習

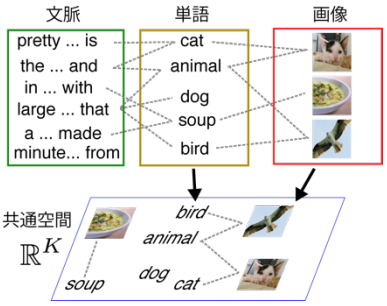
キーワード：多変量解析　次元削減　分散表現　表現学習　グラフ埋め込み　自然言語処理　ニューラルネットワーク　加法構成性

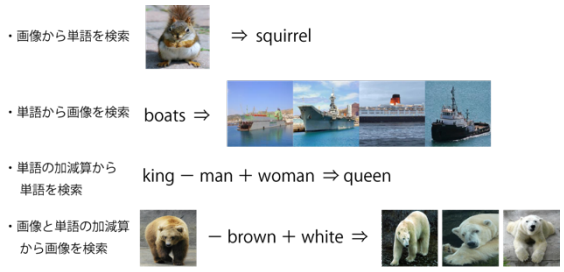
様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

画像，タグ，文書等の様々なドメインのマルチモーダルデータから関連性を考慮したグラフ埋め込みによって情報統合し共通空間で表現する方法，すなわち「多ドメイン関連性データのグラフ埋め込みによる表現学習」は，従来の多変量解析の一般化とも言える．このような多変量解析手法として，我々はクロスドメインマッチング相関分析（Cross-Domain Matching Correlation Analysis；CDMCA）を提案している（Shimodaira 2016；*Neural Networks*, Vol. 75, 126-140)．そして CDMCA の線形変換をニューラルネットによる非線形変換に置き換えて，グラフ埋め込みによる次元削減を一般化する方法を Probabilistic Multi-view Graph Embedding（PMvGE)として提案した(Okuno, Hada, Shimodaira 2018；*Proc. of the 35th International Conference on Machine Learning, ICML,* PMLR 80, 3888-3897)．ただし確率的な生成モデルを導入し，リンク重みを確率変数とみなしてモデル化した．PMvGE は，CDMCA だけでなく言語モデルによる単語ベクトルなど様々なモデルを特殊な場合として含むモデルになっている．また，二項関係をグラフで表現するかわりに，多項関係をハイパーグラフで表現する拡張を行った(Okuno and Shimodaira 2020；Neural Networks, Vol. 26, 362-383)．

これらの柔軟な多変量解析の方法論では，特徴量を線形変換やニューラルネットによる非線形変換によって次元削減し，グラフ等の構造を保存するように埋め込みの最適化をおこなう．考え方としてはシンプルであるが，現実の単語や画像データに適用してみると有効に機能することがわかった．例えば，多言語単語埋め込み（Oshikiri et al., 2016；ACL）や画像と単語の同時埋め込み（Fukui et al., 2016；ICIP）に CDMCA を適用してその有効性を確認した．Fukui et al.（2017；TextGraphs）では，写真共有サイト Flickr のタグ付けされた画像データから「画像―タグ」のリンク，Wikipedia の文書データから「単語―文脈」のリンクを入手し，単語ベクトルと画像ベクトルを 500 次元の共通空間に埋め込んで，単語と画像の加減算を含む相互検索を実現した．



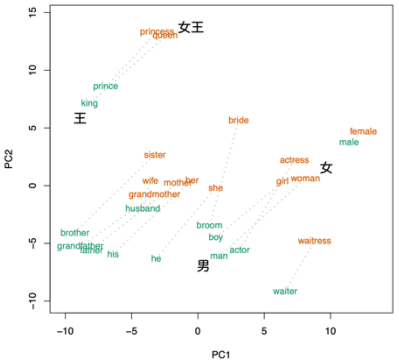


２．研究の目的

関連性データの埋め込みが表現学習として有効に機能することがわかったが，なぜこのような埋め込みが良いのか，「良い表現」とは何か？といった疑問が残る．これはニューラルネットがどのように情報を表現しているのかを理解するための一つの段階でもある．本研究では，とくに埋め込んだベクトルの加減算に代表されるような共通空間の演算の仕組みや，それに関連した埋め込み表現の性質を明らかにすることを目的とする．

３．研究の方法

単語埋め込み（単語ベクトルともいう）の加減算によってアナロジーを計算できることは，一般によく知られている．右図は，Wikipedia の文書データから事前学習済みの 300 次元の単語埋め込み GloVe から単語をいくつか選び，PCA で２次元に可視化したものである．男性に関する緑色の単語と，女性に関する朱色の単語が平行して並んでいる．特に男―女の線分と王―女王の線分が並行で長さがほぼ等しいことから，（女ベクトル － 男ベクトル）＋ 王ベクトル ＝ 女王ベクトル のようなベクトルの加減算が成立する．このような演算が可能となる背景に，加法構成性

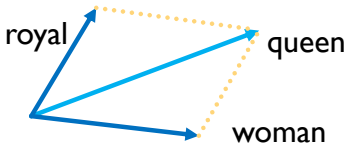


と呼ばれる性質がある．「王 = 男＋王的なもの」,「女王＝女＋王的なもの」と意味の構成要素の和に分解できると仮定すれば,「王的なもの」を２つの式から消去すると，女王＝女＋（王―男）＝（女―男）＋王が導出できる．したがって，演算の仕組みを明らかにするには，加法構成性や，それに関連する性質を調べれば良い．

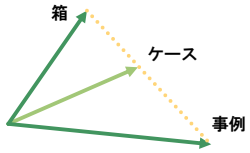
４．研究成果

主な成果について概要を述べる．

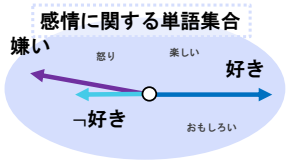
（１）通常の加法構成性では，構成要素のベクトルの和によって意味が合成される．このとき，構成要素が表す意味が同時に成立，つまり AND を表す（例: 女王 = 女 and 王的なもの）．それでは，構成要素が表す意味のどちらかが成立，つまり OR を表す(例: ケース = 箱 or 事例)にはどうすればよいか？ 埋め込みが一種の条件付き確率モデルで定義されることから，AND は確率の積，OR は確率の和と考えると，この回答が得られる．その結果，OR は構成要素の出現頻度で重み付けした重心を計算すればよいことを導出した．また, NOT はどうだろうか？ これには，まず考慮の対象となる単語または意味の構成要素の集合を考える必要がある．このような集合における埋め込みの重心を原点に取り直すと，NOT は負の方向であることが導出できる．



通常の加法構成性（AND）



頻度重み重心（OR）



負の方向（NOT）

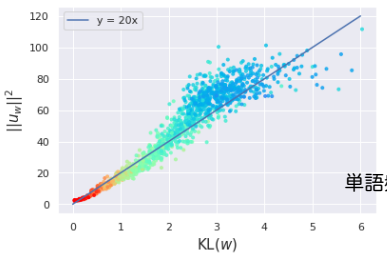
（２）２つの埋め込みの近さをベクトルのコサイン類似度で測ると，一般にタスクにおける性能が高くなることが経験的に知られている．すなわち，ベクトルの方向が意味を表していると考えられる．それではベクトルの長さは何を表しているか？ Skip-Gram with Negative Sampling (SGNS)と呼ばれる一種の対照学習によって得られる単語ベクトルについて詳しく調べた．この場合，文脈における単語の出現確率を指数型分布族で表すことができて，理論解析が可能になる．その結果，ベクトルの長さは「意味の強さ」を表すことがわかった．より正確には，単語wのベクトルu_wを適切に白色化すると，そのノルムの２乗がカルバック・ライブラー(KL)情報量

$$\mathrm{KL}\big(p(\cdot\,|w)\,\|\,p(\cdot)\big) = \sum_{w' \in V} p(w'|w)\log\frac{p(w'|w)}{p(w)}$$

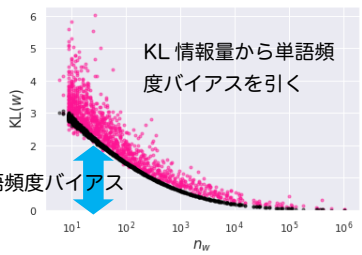
$p(\cdot\,|w)$ 単語 w の周辺単語分布

$p(\cdot)$ ユニグラム分布

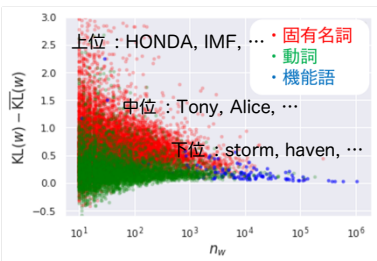
で近似できる．ベイズの立場で言えば，単語頻度の事前分布$p(\cdot)$と単語wを観測したときの事後分布$p(\cdot\,|w)$がどれだけ異なるかを KL 情報量で測ったものである．理論だけでなく実験で確かめると，かならずしも白色化しなくても(下の左図)のようにノルムの２乗が KL 情報量に比例する傾向がある．KL 情報量から注意深く頻度バイアスを引くと（中図），実際に単語の「意味の強さ」を表す傾向があった（右図）．中図と右図の横軸n_wは単語頻度である．



白色化しない埋め込みでもノルムの２乗は KL 情報量に比例



単語 w の KL 情報量（赤）とシャッフルしたコーパスの KL 情報量（黒）



バイアス補正済みの KL 情報量が，意味の強さを表す

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　3件）

１．著者名 Inoue Masaaki、Pham Thong、Shimodaira Hidetoshi	４．巻 10
２．論文標題 A Hypergraph Approach for Estimating Growth Mechanisms of Complex Networks	５．発行年 2022年
３．雑誌名 IEEE Access	６．最初と最後の頁 35012～35025
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2022.3143612	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Pham Thong、Sheridan Paul、Shimodaira Hidetoshi	４．巻 9
２．論文標題 Non-parametric estimation of the preferential attachment function from one network snapshot	５．発行年 2021年
３．雑誌名 Journal of Complex Networks	６．最初と最後の頁 cnab024
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/comnet/cnab024	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Akifumi Okuno, Hidetoshi Shimodaira	４．巻 33
２．論文標題 Extrapolation Towards Imaginary 0-Nearest Neighbour and Its Improved Convergence Rate	５．発行年 2020年
３．雑誌名 Advances in Neural Information Processing Systems (NeurIPS 2020)	６．最初と最後の頁 21889-21899
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計11件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 Masahiro Naito ,　Sho Yokoi ,　Geewook Kim ,　Hidetoshi Shimodaira
２．発表標題 Revisiting Additive Compositionality: AND, OR and NOT Operations with Word Embeddings
３．学会等名 ACL-IJCNLP 2021 Student Research Workshop（国際学会）
４．発表年 2021年

１．発表者名
大山百々勢，横井祥，下平英寿

２．発表標題
単語ベクトルの長さは意味の強さを表す

３．学会等名
言語処理学会第28回年次大会(NLP2022)

４．発表年
2022年

１．発表者名
CAO Ruixing ， 田中卓磨 ， 奥野彰文 ， 下平英寿

２．発表標題
マルチスケールk-近傍法における回帰関数および損失関数の検討

３．学会等名
第35回人工知能学会全国大会（JSAI2021）

４．発表年
2021年

１．発表者名
田中卓磨 ， 奥野彰文 ， 下平英寿

２．発表標題
マルチスケールk-近傍法による画像のExtreme Multi-Label分類

３．学会等名
第35回人工知能学会全国大会（JSAI2021）

４．発表年
2021年

１．発表者名
操瑞行 ， 田中卓磨 ， 奥野彰文 ， 下平英寿

２．発表標題
マルチスケールk-近傍法における外挿モデルの検討

３．学会等名
2021年度統計関連学会連合大会

４．発表年
2021年

１．発表者名
内藤雅博 ， 横井祥 ， 下平英寿

２．発表標題
単語埋め込みの加法構成性の精緻化と論理演算

３．学会等名
2021年度統計関連学会連合大会

４．発表年
2021年

１．発表者名
井上雅章 , THONG Pham ， 下平英寿

２．発表標題
任意のノード特徴量による成長機構をもつハイパーグラフモデル

３．学会等名
2021年度統計関連学会連合大会

４．発表年
2021年

１．発表者名
THONG Pham ， PAUL Sheridan ， 下平英寿

２．発表標題
複雑ネットワークの成長過程を観測できない時の優先的選択関数の推定方法

３．学会等名
2021年度統計関連学会連合大会

４．発表年
2021年

１．発表者名
奥野 彰文, 下平 英寿

２．発表標題
仮想的なゼロ近傍への外挿とその収束レートについて

３．学会等名
2020年度統計関連学会連合

４．発表年
2020年

１．発表者名 内藤雅博，横井祥，下平英寿
２．発表標題 単語埋め込みによる論理演算
３．学会等名 言語処理学会第27回年次大会(NLP2021)
４．発表年 2021年

１．発表者名 横井祥，下平英寿
２．発表標題 単語埋め込みの確率的等方化
３．学会等名 言語処理学会第27回年次大会(NLP2021)
４．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関