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　２ 年　 ６ 月 １２ 日現在

機関番号：３３９２４
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００３１８

研究課題名（和文）分野オントロジと大量文書に対する統合埋め込みベクトル構築

研究課題名（英文）Unified word embeddings for domain Ontologies and large-scale documents

研究代表者
佐々木　裕（Yutaka, Sasaki）

豊田工業大学・工学（系）研究科（研究院）・教授

研究者番号：６０３９５０１９

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：単語や文を深層学習の入力データとして利用するためには，単語や文の意味を数値ベクトルとして表現する必要がある．本研究では，異質な2種類の情報源である文書情報と知識ベースを，Poincare埋め込み技術をベースとして統合埋め込みベクトル空間に写像するレトロフィッティング法を開発した．実験により，文書から構築したPoincare GloVeベクトルを，概念の上下関係から作成したPoincare 埋め込みベクトルを骨格として再配置することで，単語の上下関係の予測性能を大きく改善できることを明らかにした．また，知識ベースと文書を関連付けるため，交通オントロジーに基づく新しい交通コーパスを構築した．

研究成果の学術的意義や社会的意義
単語や文を数値ベクトルとして表現する技術は深層学習において非常に重要である．しかし，これまで単語を数値ベクトルとして表現する技術と知識構造を埋め込みベクトルとして表現する技術はそれぞれ独立に研究されてきた．特に，Poincare埋め込みは単語の上位下位関係の表現に適しており，知識ベースのベクトル表現に非常に適した方法であるが，扱える単語に制約があった．本研究では，レトロフィッティング法により，大量文書から作成したPoincareベクトルを概念構造から作成したPorincare埋め込みに適合させた点に意義がある．また，交通に関する知識と文書をリンクした新コーパスを一般公開し社会に貢献する．

研究成果の概要（英文）：For applying deep neural learning, it is crucial to represent words and sentences as numerical vectors that reflect the original semantic contents. In this study, we developed a new method to integrate documents information and knowledge structures into a single space on the basis of the Poincare embedding technology. Our retrofitting method maps textual Poincare GloVe embedding vectors to the hypernym Poincare embedding space so that the similarity among textual embedding vectors are preserved. Experimental results show that we can improve hypernym detection performances of the textural Poincare embedding vectors by retrofitting the textual vectors to the skeleton of hypernym embedding vectors. In addition, we created a traffic-rule corpus that has links from traffic terms to a traffic Ontology concepts and relations.

研究分野：自然言語処理

キーワード：Poincare埋め込み　Poincare Glove　レトロフィッティング　オントロジー埋め込み　交通知識ベース

１．研究開始当初の背景

従来，文情報とオントロジを対応付けるには，文情報をSPARQL クエリに変換し，SPARQL により照合する方法が採用されてきた．このアプローチは，専門分野の表現法が限定される場合には変換パターンが限られるため有効であるが，自然言語による自由な表現を持つ文を，オントロジと整合するSPARQLクエリに変換することは非常に難しい．たとえば，「広い道路の制限速度は 60km」と「<一般道>-<法定速度>-<60km/h>」というオントロジ中の属性関係は直接照合できない．従来は，「広い道路」「制限速度」「60km」を固有表現抽出により，それぞれ「<一般道>」「<法定速度>」「<60km/h>」という概念クラスに対応づけてきたが，オントロジに大量に階層的な概念クラスが存在する場合，様々な用語表現に対応する十分な用語学習データを準備することは困難である．

そこで，単語の埋め込みベクトルにより単語間の類似度を求めるのと同様に，単語とオントロジクラス間の類似度，および単語関係とオントロジの属性関係の類似度を，共通の埋め込みベクトル空間で対応づける方法が求められている．先の例では，「広い道路～」を表す述語・項構造のベクトルとオントロジ上の「<一般道>-<制限速度>-<60km/h>」のベクトルを比較し，文の内容をオントロジと柔軟に照合できることが望まれている．

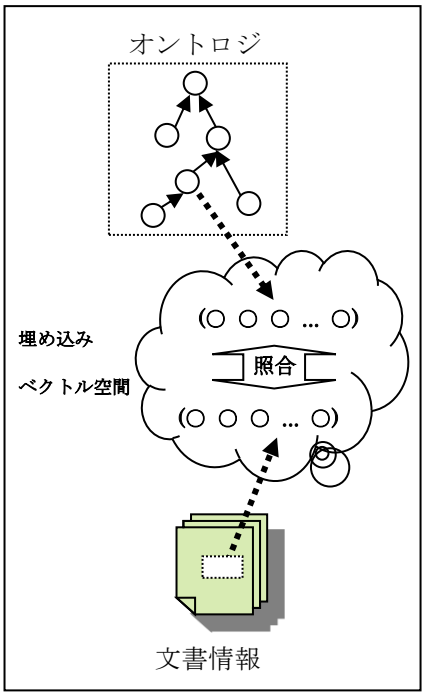

図１　提案手法

このような背景のもとで，埋め込みベクトルに記号的構造情報を導入する研究がいくつか行われている．提案者らは，2015 年に，シソーラスの持つ対義語関係(Antonymy)を扱える埋め込みベクトル表現を計算する半教師あり学習法を開発し，国際会議 NAACL-2015 で発表している [1]．2016 年には，シソーラスの持つ上位・下位語関係(Hypernymy)に基づき，下位語の埋め込みベクトルを上位語の埋め込みベクトルに変換する写像の学習法を開発し，国際会議 COLING-2016 で発表した[2]．さらに，上位・下位関係に関しては，2013～2015 年にかけて科研費基盤(C)により，大規模階層的文書分類を高速・高精度に実行する手法を開発しており，LSHTC Wikipedia データにおいて世界最高の精度を実現した．

従来，単語の埋め込み表現と知識の埋め込み表現は独立して研究されてきた．単語の埋め込み表現に関しては，word2vec 等の埋め込み表現技術が近年の言語処理において必須の基盤技術となっている．本研究の提案後に，概念の上下関係に関しては，2017 年に Nickel and Kiela[3]により，双曲空間上での Poincaré 埋め込み表現[4]（図 2）が提案され非常に高い性能が報告されたため，本研究の軌道修正を行った．この埋め込み表現は空間の中心から周辺に向かった空間が広がっていくため，階層構造の表現に非常に適している．文書中の単語の GloVe 埋め込みを双曲空間で表現した Poincaré GloVe(図 3)も 2018 年に提案されたが，両者は独立しており，オントロジの subClassOf 関係（上位・下位関係）と文書に現れる単語の埋め込み表現は別の空間で表現されていた．Poincaré 埋め込みは，辞書から取り出した少量の概念の上下関係から構築されるため限られた単語の範囲では非常に高精度に上下関係が予測できるが，多様な単語には対応できないという欠点がある．一方，Poincaré GloVe は対象の文書から構築され，十分な単語のカバー率があるが，単語の上下関係を骨格として表現できていないという問題点が存在する．

そこで，本研究では，Poincaré 埋め込みにより構築された骨格となる単語の上下関係の埋め込みベクトルに対して，Poincaré Glove により大量文書から構築された埋め込みベクトルを再配置するというアプローチを考案した．これをレトロフィッティングと呼ぶ．また，これらの研究と並行して，評価用のデータセットとして利用するため，交通文書に対するアノテーションを実施した．

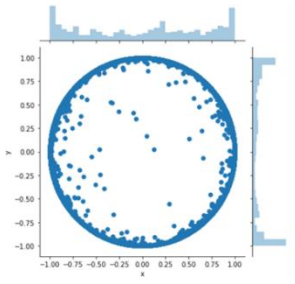

図２　Poincaré 埋め込み（疎）
（中心から周辺に単語を階層化）

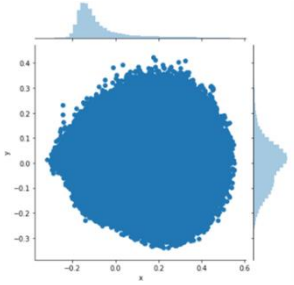

図３　Poincaré Glove 埋め込み（密）
（大量の単語が密に配置される）

２．研究の目的

単語や文章を数値ベクトル表現に変換する手法は，近年，盛んに研究されており，急速に発展している．word2vec や GloVe に続いて 2018 年に開発された BERT は，現在，文脈を考慮した単語の埋め込み表現を得る方法として標準的に使われており，深層学習のモデルを変更することなく，入力の単語の埋め込み表現を BERT に置き換えるだけで，様々な NLP アプリケーションにおいて最高性能が報告されている．

本研究では，このような文脈を考慮した埋め込みではなく，独立した単語の埋め込みベクトル表現により，単語間の上下関係と類似性を同一空間内で表現することを目的とする．これにより，大量の単語に対して，上下関係を反映した埋め込みベクトルを得ることができる．

また，関連する研究として，単語間の上下関係の推定を分類問題として学習したり，上下関係の変換を学習する研究（TransE 等）が存在するが，ここでは，変換法を学習したり，上下関係を分類問題として解いたりするのではなく，すべての単語の埋め込みベクトルのなかに自然な形で上下関係を埋め込むことを目指している．つまり，単語ベクトル間の単純な比較で上下関係と類似性の両方を得ることができる．このことにより，深層学習モデルの入力として単語を与えるときに，その表現ベクトルに自然に上下関係が含まれるというメリットがある．ただし，アプリケーションでの検証は本研究の対象外であり，次の研究ステップで実施する．本研究では，大規模な単語集合に対して，類似性と上下関係を同一空間の Poincaré 埋め込み表現により高い品質で表すことを目的とする．

３．研究の方法

全体の流れを図 4 に示す．

(1) Poincaré 埋め込み

Nickel ら[3]によって提案された Poincaré 埋め込みでは，WordNet に定義された単語の上位下位関係を元に，サンプリング手法を用いて双曲空間に単語をベクトル表現する．作成される単語表現は，空間の中心に近づくほど上位の概念になっていく性質（図 4①）を持つため，ベクトルの座標情報から上位下位関係を推定することが出来る．

(2) Poincaré GloVe

Tifrea ら[4]によって提案された Poincaré GloVe では Wikipedia のラベルなしテキスト情報を元に，重み付き最小二乗法を用いて双曲空間に単語をベクトル表現する（図 4②）．作成される単語表現の分布は上位下位関係と相関があり，単語の上位下位関係を分布情報から推定することが出来ると報告されている．

(3) Retrofitting

Faruqui ら[5]によって提案された Retrofitting では， 少量の質の良いデータを用いて大規模な単語ベクトルの改善を行う方法である．ただし，WordNet の構造そのものに単語埋め込みを再配置する研究であり，本研究のように Poincaré 埋め込みと Poincaré GloVe を対象にしたものではない．

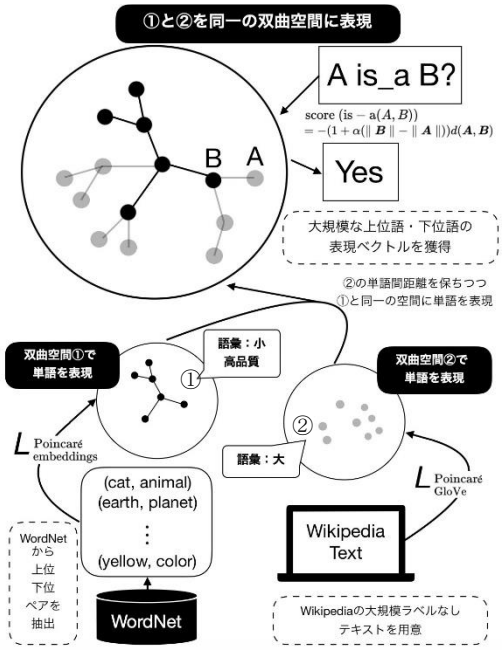

図 4 手法の流れ

(4) 変換手法

図 4 のように，双曲距離に基づいて，Poincaré GloVe のベクトルをその距離を保つように Poincaré 埋め込み空間上に配置・更新し，作成したベクトル表現の座標情報から上位下位スコアを計算・評価する．まず，Poincaré 埋め込みと Poincaré GloVe の両手法を用いて，それぞれのベクトル表現集合①と②を作成し，これらのベクトル表現を元に手法の前提となるベクトル（基盤ベクトル）を作成する．次に，目的関数を最小化するようにベクトルを更新することで，Poincaré GloVe ベクトルを Poincaré 埋め込みのベクトル空間で表現する．最後に，作成したベクトルを評価用データセットにより評価する．

４．研究成果

実験は，まず，事前実験として Poincaré 埋め込みと Poincaré GloVe のそれぞれのベクトル分布の考察を行なった後に，両手法のベクトル表現を用いて提案手法の評価を行なった．

(1) 実験設定

提案手法の評価では，Wikipedia から作成した語彙数 208,881 の Poincaré GloVe ベクトルと，WordNet [7]から作成した語彙数 31,222 の Poincaré 埋め込みベクトルを用いて大域的更新および局所的更新の場合の最適化を行い，評価したときの Baroni2012[7]，WBLESS[8] データセットでの正解率を比較した．ベクトルの次元は 100，最適化時のエポック数は 300，学習率は大域的更新で 10^{-6}，局所的更新で 10^{-4} とした．

(2) ベクトル分布の比較

Poincaré 埋め込みと Poincaré GloVe の各手法により作成されるベクトル分布の差異を説明する．次元双曲空間にベクトル表現した例は図 2 と図 3 に示した． Poincaré 埋め込みでは空間の周縁部に多くのベクトルが表現されるのに対して，Poincaré GloVe では空間の中心部に多くのベクトルが表現されており，ベクトル分布に大きな違いがある．この違いにより両手法では異なるベクトルの評価方法が用いられているが，本研究では Poincaré GloVe ベクトルを更新して Poincaré 埋め込みの空間に表現し，単一の評価方法による評価を行う．

(3) 提案手法の評価

目的関数を大域的に更新する方法と局所的に更新する方法のそれぞれを用いてベクトルを更新し，得られたベクトル表現を Baroni2012 および WBLESS 評価データセットを用いて評価した結果を表 1 に示す．レトロフィッティング前の Poincaré GloVe ベクトルの正解率はそれぞれ約 0.54 と約 0.7 であった．結果は，いずれの評価データセットにおいても局所的更新の場合の方が高い正解率であった．これは，局所的に更新を行う方が少ない制約のもと自由にパラメータを更新できるためではないかと考えられる．

	Baroni2012	WBLESS
収録単語ペア数	2,770	1,668
正解ペア数（大域的更新）	2,000	1,279
正解率（大域的更新）	0.722	0.767
正解ペア数（局所的更新）	2,102	1,329
正解率（局所的更新）	**0.759**	**0.797**

表１　評価データセットによる評価結果

(4) 交通文書へのアノテーション

オントロジと文書の間に明確なリンクを張ることで文書情報とオントロジの間に意味的な関係を明示することができる．本研究では，交通オントロジを交通教則文にアノテーションすることにより，本研究室で構築している交通オントロジとリンクさせた（例：図 5）．本コーパスは，細部の調整を実施したあと，一般に公開していく．

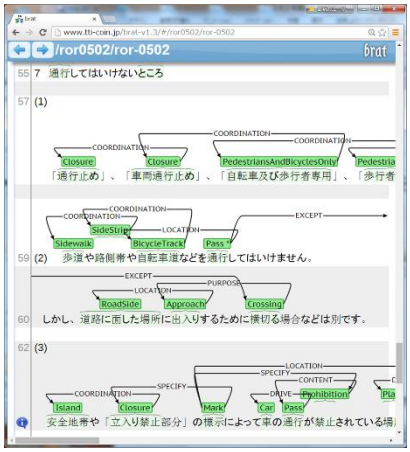

図 5 オントロジアノテーションの例

参考文献

[1] Masataka Ono, Makoto Miwa, and Yutaka Sasaki, Word Embedding-based Antonym Detection using Thesauri and Distributional Information, 2015 Conference of the North American Chapter of the Association for Computational Linguistics / Human Language Technologies (NAACL/HLT-2015), pp. 984-989, Denver, USA, June 2015.

[2] Josuke Yamane, Tomoya Takatani, Hitoshi Yamada, Makoto Miwa, and Yutaka Sasaki, Distributional Hypernym Generation by Jointly Learning Clusters and Projections, 26th International Conference on Computational Linguistics (COLING-2016), Osaka, December 2016.

[3] Maximilian Nickel and Douwe Kiela, Poincaré embeddings for learning hierarchical representations, Advances in Neural Information Processing Systems, pp. 6338-6347, 2017.

[4] Alexandru Tifrea, Gary Becigneul, and Octavian-Eugen Ganea, Poincaré GloVe: Hyperbolic word embeddings, International Conference on Learning Representations, 2019.

[5] Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith, Retrofitting word vectors to semantic lexicons, 2015 Conference of the North American Chapter of the Association for Computational Linguistics / Human Language Technologies, pp. 1606-1615, 2015.

[6] George A. Miller, WordNet: A lexical database for English, Communications of the ACM, Vol. 38(11), pp. 39-41, 1995.

[7] Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-Chieh Shan, Entailment above the word level in distributional semantics, 13th Conference of European Chapter of the Association for Computational Linguistics, pp. 23-32, 2012.

[8] Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller, Learning to distinguish hypernyms and co-hyponyms, 25th International Conference on Computational Linguistics: Technical Papers, pp. 2249-2259, 2014.

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計6件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 Marc Evrard, Makoto Miwa, Yutaka Sasaki
２．発表標題 Semantic Graph Embeddings and a Neural Language Model for Word Sense Disambiguation
３．学会等名 Second International Workshop on Symbolic-Neural Learning (SNL2018)（国際学会）
４．発表年 2018年

１．発表者名 Marc Evrard, Makoto Miwa, Yutaka Sasaki
２．発表標題 TTI's Approaches to Symbolic-Neural Learning
３．学会等名 International Workshop on Symbolic-Neural Learning（国際学会）
４．発表年 2017年

１．発表者名 Savong Bou, Naoki Suzuki, Makoto Miwa, Yutaka Sasaki
２．発表標題 Ontolgy-Style Relation Annotation: A Case Study
３．学会等名 LREC 2020 (accepted)（国際学会）
４．発表年 2020年

１．発表者名 村瀬敦也，三輪誠，佐々木裕
２．発表標題 Poincae GloVe ベクトルのレトロフィッティング
３．学会等名 言語処理学会第26回年次大会
４．発表年 2020年

１．発表者名 Savong Bou，鈴木直樹，三輪誠，佐々木裕
２．発表標題 オントロジー形式による交通関係アノテーション
３．学会等名 言語処理学会第26回年次大会
４．発表年 2020年

１．発表者名 鈴木直樹，Savong Bou，三輪誠，佐々木裕
２．発表標題 オントロジー形式アノテーションを対象とした交通用語・関係抽出と正誤問題の回答
３．学会等名 言語処理学会第26回年次大会
４．発表年 2020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
研究協力者	ＥＶＲＡＲＤ　Ｍａｒｃ (EVRARD Marc)	豊田工業大学・PD研究員 (33924)	
研究協力者	ＢＯＵ　Ｓａｖｏｎｇ (BOU Savong)	豊田工業大学・PD研究員 (33924)	
研究協力者	鈴木　直樹 (SUZUKI Naoki)	豊田工業大学 (33924)	

６．研究組織（つづき）

	氏名 （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	村瀬　敦也 (MURASE Atsuya) (33924)	豊田工業大学	
連携研究者	三輪　誠 (MIWA Makoto) (00529646)	豊田工業大学・工学部・准教授 (33924)	