# 擬ユークリッド空間への単語埋め込み

Geewook Kim[†,‡]　奥野 彰文[†,‡]　下平 英寿[†,‡]
† 京都大学大学院 情報学研究科　‡ 理化学研究所 革新知能統合研究センター
{geewook, okuno}@sys.i.kyoto-u.ac.jp, shimo@i.kyoto-u.ac.jp

## 1 はじめに

単語の意味や単語間の関係性を機械に理解させることは様々な自然言語処理において重要な課題である．単語埋め込みとは，単語の意味や単語間の関係性を内包する単語ベクトル（分散表現）を獲得する手法のことであり，単語ベクトルの利用先はニューラルネットの初期値として利用する方法，単語の意味的な距離尺度として利用する方法など，多岐にわたる．

多くの単語埋め込みは，Wikipedia のようなテキストコーパスや WordNet (Fellbaum, 1998) のような言語資源からの単語間の関係情報（例えばコーパス内の単語ペアの共起回数）を単語の分散表現を用いてモデリングすることで学習を行う．このような単語埋め込みでは一般に単語ベクトル間の内積を用いて単語間の類似度をモデリングする (Mikolov et al., 2013; Bojanowski et al., 2017; Okuno et al., 2018a)．しかし近年，通常の内積とは異なる様々な類似度関数が表現学習において試されており，その有用性が研究されている (Nickel and Kiela, 2017, 2018; Leimeister and Wilson, 2018)．特に擬ユークリッド空間への埋め込みは通常の内積による埋め込みはもちろん双曲空間への埋め込みなどを表現能力の意味で内包しており，高い表現能力を持つことが証明されている (Okuno et al., 2018b)．

このような表現学習の成果をふまえ，本稿では擬ユークリッド空間への単語埋め込みを提案する．提案手法は豊富な表現能力を持つため，既存手法では埋め込めなかった単語間の関係情報を内包する単語の分散表現を得ることが期待される．提案手法は，既存の埋め込み手法が通常の内積を用いて分散表現間の類似度を表現していたところを，擬ユークリッド内積へ置き換えることで簡単に実現できる．実験では提案手法の有用性を示すために，Wikipedia テキストコーパスと WordNet を用いた単語埋め込みを行い，それぞれモデルの表現力に対する評価と単語ベクトルの質に対する評価を行った．

## 2 提案手法

本手法では，単語ベクトル間の類似度により単語間の関係性を再現できる形で単語をベクトル空間へ埋め込むことを考える．提案モデルは既存の多くの表現学習手法が分散表現の内積により類似度を定義していたところを，擬ユークリッド内積へ置き換えることで簡単に実現できる．

### 2.1 確率グラフモデル

まず，表現学習で広く用いられている以下の確率グラフモデルを考える (Mikolov et al., 2013; Tang et al., 2015)．大量のテキストコーパスや WordNet のような言語資源から単語間の関係情報が与えられており，その情報が重み付き無向グラフ $G = (V, E)$ で表されているとする．ここで，ノード (単語) 集合は $V$，エッジ集合は $E$ である．そしてグラフ上でノード $i$ と $j$ の同時確率，すなわち，エッジが存在する確率を以下のようにモデリングする．

$$p(i', j') = \frac{\exp\left(s(x_{i'}, x_{j'})\right)}{\sum_{1 \le k \le l \le |V|} \exp\left(s(x_k, x_l)\right)}, \qquad (1)$$

ただし，$i' = \min(i, j), j' = \max(i, j)$ であり，$x_i$ は $i$ 番目の単語の分散表現，そして $s(\cdot, \cdot)$ は分散表現間の類似度を表す関数であり，一般に内積が用いられる．

### 2.2 提案モデル

2.1 節の確率モデルにおいて，提案モデルは単語ベクトル間の類似度を擬ユークリッド内積によって定義することで実現する．

**定義 1** 通常の内積を $\langle\cdot,\cdot\rangle_{\mathbf{E}}$，ユークリッド空間を $\mathbf{E}^n$ で表し，$a = a_1^{m+n}, b = b_1^{m+n} \in \mathbb{R}^{m+n}$ とする．擬ユークリッド空間 $\mathbf{P}^{m,n}$ は，直和集合 $\mathbf{E}^m \oplus \mathbf{E}^n$ と擬ユークリッド内積 $\langle a,b\rangle_{\mathbf{P}} = \langle a_1^m, b_1^m\rangle_{\mathbf{E}} - \langle a_{m+1}^{m+n}, b_{m+1}^{m+n}\rangle_{\mathbf{E}}$ の組で定義される空間である．

したがって，$i,j$ 番目の単語に対応する分散表現 $x_i, x_j$ 間の類似度を

$$s(x_i, x_j) = \langle x_i, x_j\rangle_{\mathbf{P}}, \tag{2}$$

と定義することで提案モデルが実現できる．

### 2.3 学習

2.1 節の確率モデルの単純な学習方法として，以下のような経験分布

$$\hat{p}(i,j) = \frac{w_{i,j}}{\sum_{1\le k\le l\le |V|} w_{k,l}}, \tag{3}$$

とモデル $p(\cdot,\cdot)$ との KL ダイバージェンスを最小化する方法が提案されている (Tang et al., 2015)．これは以下のような式を最小化することに対応するが，

$$\sum_{(i,j)\in E} \hat{p}(i,j)\log\frac{\hat{p}(i,j)}{p(i,j)}, \tag{4}$$

上式から定数項を無視することで，次の目的関数 $G(X)$ を得ることができる．

$$G(X) = \sum_{(i,j)\in E} w_{i,j}\log p(i,j). \tag{5}$$

式 (5) を単語 $i$ の分散表現 $x_i$ に対して偏微分すると，

$$\frac{\partial G(X)}{\partial x_i} = \sum_{j\in\{j|(i,j)\in E\}} w_{i,j}\frac{\partial \log p(i,j)}{\partial x_i}, \tag{6}$$

となるが，Tang et al. (2015) と同様，実際に勾配降下法を行う時には重み $w_{i,j}$ のエッジ $(i,j)$ を重み 1 の $w_{i,j}$ 個の複数のエッジとみなしてよい．また，式 1 からわかるようにノードの数に比例してモデルの計算コスト膨大になってしまう問題があるため，Mikolov et al. (2013) で提案されている Negative sampling 手法も併用した．具体的には，エッジ $(i,j)$ に対して，以下のような目的関数を最大化するように学習を行った．

$$\log\sigma(\langle x_i, x_j\rangle_{\mathbf{P}}) + \sum_{n\sim P_{\text{neg}}} \log\sigma(-\langle x_{i'}, x_n\rangle_{\mathbf{P}}), \tag{7}$$

ここで $P_{\text{neg}}$ はネガティブサンプリングのためのノードに対する分布であり，本稿ではノードの頻度に比例する経験分布を用いた．また，$i'$ は $\{i,j\}$ から一様にランダムサンプリングした．

### 2.4 表現学習の既存手法との対応関係

表現学習において擬ユークリッド空間への埋め込みは Okuno et al. (2018b) で提案されている．これと他の既存手法との対応関係は以下のようになる．まず，$n = 0$ にすると通常のユークリッド空間への埋め込みとなる．次に，負の空間の次元 $n$ を 1 とすることで，埋め込み空間はミンコフスキー空間となるが，ここでさらにノルムを $-1$ へ制約する条件を入れると，Nickel and Kiela (2017) や Leimeister and Wilson (2018) や Nickel and Kiela (2018) で用いる双曲空間への埋め込みに対応し，これはポアンカレ埋め込みとも呼ばれる．最後に，Okuno et al. (2018b) で提案されている定数項付き内積 (Shifted Inner Product) による表現学習法は表現能力の意味で擬ユークリッド空間への埋め込みに内包されることが証明されている (Okuno et al., 2018b, E.1 章)．

## 3 実験

提案手法を評価するために 2 つのタスクを行った．まず，WordNet の単語の階層情報を用いて単語の分散表現を学習して，得られた分散表現から元の階層グラフを予測するタスクを行い，モデルの表現能力を評価した．次に，テキストコーパス上の単語間の共起情報を用いて単語の分散表現を学習して，様々なベンチマークデータセットを用いて単語類似度タスクを行い，分散表現の質を評価した．

### 3.1 グラフ再構築タスク

本タスクでは入力として与えられたグラフを用いて分散表現を学習したあと，得られた分散表現からノード間のエッジの有無を予測するタスクを行う．つまり，元のグラフ構造を予測するタスクを行い，その性能を ROC-AUC (Bradley, 1997) より評価した．実験の設定は Okuno et al. (2018b) とほぼ同じである．実験ではパラメーター数の影響をみるために，単語ごとに割り当てるパラメーターの総数 $K$ を $\{10, 20, 50, 100\}$ と変化させて実験を行った．

**データセット**: WordNet の単語の階層ツリーを用いた．データセットの前処理は Nickel and Kiela (2017) の実験と同様であるが，単語の多義性を表す追加素性は本実験では利用しなかった．その結果，単語数は 67,186 であり，エッジ数は 682,169 である．

表 1: 階層構造再構築タスクの実験結果（ROC-AUC [%]）

| モデル | K=10 | K=20 | K=50 | K=100 |
|---|---|---|---|---|
| LINE1st (Tang et al., 2015) | 90.50 | 91.78 | 91.96 | 91.95 |
| LINE2nd (Tang et al., 2015) | 83.79 | 83.14 | 77.91 | 77.79 |
| SIPS (Okuno et al., 2018b) | 96.34 | 96.85 | 97.03 | 96.98 |
| Poincaré Embedding (Nickel and Kiela, 2017) | 93.28 | 93.91 | 93.97 | 93.80 |
| **提案手法** | **99.68** | **99.88** | **99.90** | **99.90** |

表 2: 単語類似度タスクの実験結果（スピアマン順位相関 [%]）

| | K=10 | | | | K=100 | | | |
|---|---|---|---|---|---|---|---|---|
| モデル | MEN | WS | RG | SimLex | MEN | WS | RG | SimLex |
| Skip-gram (Mikolov et al., 2013) | 33.29 | 34.57 | **51.07** | 15.73 | 59.36 | 65.11 | 57.24 | 26.13 |
| GloVe (Pennington et al., 2014) | 19.10 | 19.38 | 11.58 | 7.63 | 41.06 | 46.56 | 36.43 | 16.18 |
| Hyperbolic Skip-gram (Leimeister and Wilson, 2018) | 41.39 | 42.25 | 44.68 | **19.65** | 49.23 | 52.45 | 45.70 | 21.30 |
| **提案手法** | **48.92** | **50.19** | 44.38 | 15.57 | **66.52** | **68.75** | **60.37** | **26.24** |

**ベースライン手法**: 内積を持いて分散表現間の類似度を表現する手法として，まずLINE1st と LINE2nd (Tang et al., 2015) を用いた．ただし LINE2nd では Skip-gram (Mikolov et al., 2013) のように，単語ごとに 2 つの分散表現を与える点に注意が必要である．例えば $K = 100$ で LINE2nd は単語に対して 50 次元の 2 つのベクトルを設ける．次に，バイアス項付きの内積によるグラフ埋め込み手法 SIPS (Okuno et al., 2018b) と，WordNet のようなツリーを埋め込むための有効性が知られている Poincaré Embedding (Nickel and Kiela, 2017) もベースラインとして用いた．

**学習**: バッチサイズ 64，総イテレーション数 50 万の確率的勾配降下法を用いて学習を行った．そして，5 万イテレーションごとに評価を行い，最高点を結果として示した．学習率の初期値 $\rho_0$ は $\{0.01, 0.02\}$ から探索し，Mikolov et al. (2013) と同様にイテレーションの総数を $T$ としたとき $t$ イテレーション目の学習率は $\rho_t = \rho_0(1-t/T)$ とした．提案手法では $\mathbf{P}^{m,n}, m+n = K$ における次元 $n$ は {1,2,4,8} より探索した．

**結果**: 実験結果は表 1 に与えた．全てのケースにおいて提案手法が一番高い AUC を示している．

## 3.2 単語類似度タスク

本タスクではテキストコーパスを用いて単語埋め込みを行い，単語間の類似度スコア（提案手法では擬ユークリッド内積）と人手で付けられた単語間の意味的類似性スコアとのスピアマン順位相関を計算することで分散表現の質を評価した．

**ベースライン手法**: まず Skip-gram (Mikolov et al., 2013) と GloVe (Pennington et al., 2014) を用いた．それぞれ内積とバイアス項付き内積を用いて分散表現を学習する手法であるが，両方とも単語ごとに 2 つのベクトルを設けて学習を進める点に注意が必要である．この 2 つの手法の評価では得られた単語ベクトルのコサイン類似度を用いた．双曲空間への埋め込みの手法としては，Leimeister and Wilson (2018) を用いた．この手法はミンコフスキ空間でノルムを $-1$ へ制限した双曲空間へ単語を埋め込む．評価ではミンコフスキー内積を類似度の尺度として用いた．

**データセット**: 前処理済みのコーパス（text8） (Mahoney, 2009) を用いた．ベンチマークデータセットとしては，MEN (Bruni et al., 2012)，WS (Finkelstein et al., 2001)，RG (Rubenstein and Goodenough, 1965)，SimLex (Hill et al., 2015) を用いた．

**学習**: バッチサイズ 1 の非同期確率的勾配降下法 (Niu et al., 2011) を用いて行った．バリデーションデータセットとしては MEN を用いた．エポックは 5 にして，学習率などの他の設定は 3.1 節と同様である．

**結果**: 実験結果は表 2 のようである．提案手法が他の手法に比べてヒューマンスコアと比較的高い相関を示していることがわかる．

## 4　まとめと今後の課題

本稿では擬ユークリッド空間への単語埋め込みを提案した．提案手法は分散表現間の類似度を擬ユークリッド内積により定義することで簡単に実装できる．実験ではグラフ再構築タスクと単語類似度タスクにおいて高いスコアを示した．今後の課題として，提案手法によって得られた分散表現が既存手法による分散表現とどのように異なる性質を持つかについて詳しく調べていきたい．また，機械翻訳や品詞推定などの，単語ベクトルの応用先での提案手法の有用性についても調べていきたい．

## 参考文献


Bojanowski, Piotr, Grave, Edouard, Joulin, Armand, and Mikolov, Tomas (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bradley, Andrew P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Bruni, Elia, Boleda, Gemma, Baroni, Marco, and Tran, Nam Khanh (2012). Distributional Semantics in Technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145. Association for Computational Linguistics.

Fellbaum, Christiane (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.

Finkelstein, Lev, Gabrilovich, Evgeniy, Matias, Yossi, Rivlin, Ehud, Solan, Zach, Wolfman, Gadi, and Ruppin, Eytan (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web*, WWW'01, pages 406–414, New York, NY, USA. ACM.

Hill, Felix, Reichart, Roi, and Korhonen, Anna (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Leimeister, Matthias and Wilson, Benjamin J. (2018). Skip-gram word embeddings in hyperbolic space. *arXiv:1809.01498*.

Mahoney, Matt (2009). Large text compression benchmark.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg, and Dean, Jeffrey (2013). Distributed Representations of Words and Phrases and Their Compositionality. In *Advances in Neural Information Processing Systems*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Nickel, Maximillian and Kiela, Douwe (2017). Poincaré Embeddings for Learning Hierarchical Representations. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 6338–6347. Curran Associates, Inc.

Nickel, Maximillian and Kiela, Douwe (2018). Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry. In Dy, Jennifer and Krause, Andreas, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3779–3788, Stockholmsmssan, Stockholm Sweden. PMLR.

Niu, Feng, Recht, Benjamin, Re, Christopher, and Wright, Stephen J. (2011). HOGWILD!: A Lock-free Approach to Parallelizing Stochastic Gradient Descent. In *Advances in Neural Information Processing Systems*, NIPS'11, pages 693–701, USA. Curran Associates Inc.

Okuno, Akifumi, Hada, Tetsuya, and Shimodaira, Hidetoshi (2018a). A probabilistic framework for multi-view feature learning with many-to-many associations via neural networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 3885–3894.

Okuno, Akifumi, Kim, Geewook, and Shimodaira, Hidetoshi (2018b). Graph Embedding with Shifted Inner Product Similarity and Its Improved Approximation Capability. *arXiv:1810.03463*.

Pennington, Jeffrey, Socher, Richard, and Manning, Christopher (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Rubenstein, Herbert and Goodenough, John (1965). Contextual correlates of synonymy. *Commun. ACM*, 8:627–633.

Tang, Jian, Qu, Meng, Wang, Mingzhe, Zhang, Ming, Yan, Jun, and Mei, Qiaozhu (2015). LINE: Large-scale Information Network Embedding. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 1067–1077, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.