Word2Box を用いた人々の移動に基づく 地域メッシュの領域表現

奥島　海 [1] 廣田 雅春 [1,2]
[1] 岡山理科大学 総合情報学部　[2] 岡山理科大学 情報理工学部
i20i022ok@ous.jp　hirota@ous.ac.jp

概要

単語の埋め込み表現は，自然言語処理の様々なタスクで用いられている重要な技術である．また，地理情報に関する様々なタスクで，位置情報から得た地域メッシュの埋め込み表現が用いられている．本研究では，地域メッシュの埋め込み表現の作成に，Word2Box を用いることで集合演算が可能な埋め込み表現を作成する．Word2Box は，単語の意味の広がりや階層関係などを表すことが可能な Box Embeddings を獲得するための教師なし学習の手法である．獲得された領域表現に対して集合演算を行い，地図上に可視化した結果を分析する．

1　はじめに

モビリティ技術の発展により，様々なデバイスから移動軌跡が収集されている．それに伴い，マーケティングや，災害，交通，観光などの分野において，人々の移動パターンを分析することの重要性が高まっている．その際に，人々の移動を表現する方法として地域メッシュを用いることがある．地域メッシュとは，緯度と経度に基づいて領域をほぼ同じ大きさの網の目に分割することで得られる小さい領域のことである．

本研究では，移動軌跡から得られる人々の地域メッシュ間の移動に基づいた地域メッシュの埋め込み表現を獲得する．測位点を地域メッシュに変換したものを単語とみなし，移動軌跡を単語列とみなすことで，地域メッシュのベクトルを作成する．地域メッシュの埋め込み表現は，位置情報の予測 [1] や，移動軌跡の補間 [2]，目的地の予測 [3]，場所の推薦 [4] などの様々な分野で利用されている．

単語の埋め込み表現は，単語をベクトルで表現するための方法として一般的であり，多くのタスクで利用されている．その際には，Word2Vec [5] や，

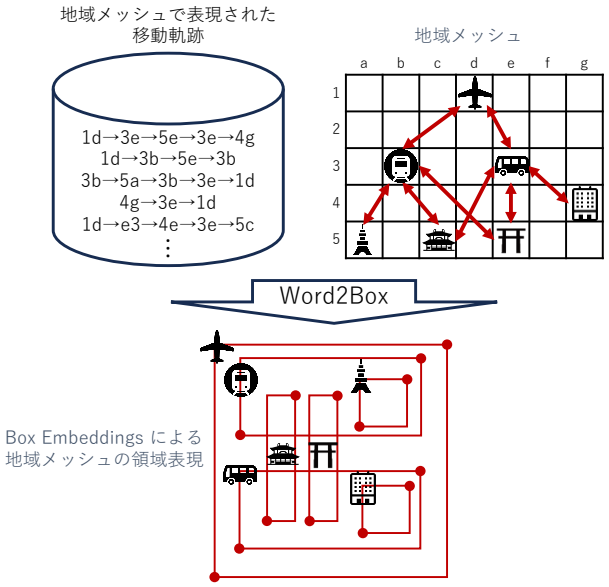


図 1: 地域メッシュ間の移動に基づいた Word2Box によって獲得される領域表現

GloVe [6]，FastText [7] などの手法が主に用いられてきた．また，これらの手法は文脈を考慮しない分散表現を獲得するが，ELMo [8]，BERT [9] などの文脈を考慮した単語の埋め込み表現を獲得可能な手法も提案されている．しかし，これらの手法では，単語をひとつのベクトルで表現するため，単語を点でしか表現することができないという課題を持つ．そのため，本研究では，単語の意味の包含関係や階層関係などの集合的な性質を表すことができる単語の Box Embeddings [10] を利用する．Box Embeddings は，単語を 2 つの点で表現することにより，単語を超直方体による領域で表現する．本研究では，教師なし学習により，Box Embeddings を獲得可能な Word2Box [11] を用いる．この手法により獲得された領域の集合演算は，tongue ∩ body は eye や mouth，tongue ∩ language は dialect や idiom が得られるなど単語の集合演算が可能である．

ここで，本研究で獲得される地域メッシュの領

域表現の例を図 1 に示す．はじめに，1d (空港) → 3d (バス停) → 5e (神社) → 3e (バス停) → 4g (ホテル) のように移動軌跡が地域メッシュの単語列で表現されているとする．このような単語列に対して，Word2Box を適用すると，図 1 の下側のような Box Embeddings による領域表現を得ることが期待される．このとき，空港からの移動に関して，移動軌跡から一定以内の移動回数によって得られる地域メッシュの種類数が最も多いならば，空港の地域メッシュの領域は最も大きくなることが期待される．これは，一般的な領域表現において，単語の意味の広がりが大きいことを表している．次に，バス停と駅は，空港との往復とホテルや神社などの往復が行われるが，それらは互いに往復での移動が行われないならば，それらの領域は空港よりも小さくなるかつ互いに重なりが存在しない領域が得られる．神社と寺は，バス停と駅のそれぞれに対して往来があるため，バス停と駅のいずれにも重なる領域が得られる．同様に，タワーとホテルは，バス停と駅のいずれかとの往復しか無いため，それぞれの領域にのみ包含された領域が得られる．本研究では，このような人々の移動に基づき，移動先の多様さを単語の意味の広がりとして扱うことで地域メッシュの領域表現を得る．

本研究では，X[1] (旧 Twitter) から得た移動軌跡を Word2Box で学習することで地域メッシュの領域表現を獲得する．また，その領域表現を用いた集合演算を行うことによる人々の移動について分析の例を示す．

2 関連研究

単語の埋め込み表現の技術は，単語以外にも文章 [12] や，絵文字 [13]，グラフのノード [14]，遺伝子 [15]，地域メッシュ [16] など様々な分野に応用されている．そのため，埋め込み表現の研究は盛んに行われており，近年では，単語の意味の広がりや階層関係を表現するために，単語を双曲空間に埋め込む手法や領域で表現する手法が研究されている．

双曲空間に単語を埋め込む手法として，ポアンカレ円盤モデルに単語を埋め込む Poincaré Embeddings [17] や，ローレンツモデルを用いた Lorentz Embeddings [18] が提案されている．また，Poincaré Embeddings を獲得するために，GloVe と組み合わせることで，教師なし学習で獲得する手法である Poincaré GloVe [19] が提案されている．

単語を領域で表現するために様々な形状や分布で埋め込む手法が提案されている．たとえば，ガウス分布 [20] や，ベータ分布 [21]，円錐 [22]，円盤 [23] で単語を表現する手法が提案されている．それらに加えて，単語の領域を超直方体で表現する Box Embeddings [10] が提案されている．その後，領域の重なりの局所的識別性を改善するために，その評価にガンベル分布を用いる Gumbel Box [24] が提案された．また，Gumbel Box を教師なし学習によって得るために Continuous Bag of Words (CBOW) を用いて学習する Word2Box [11] が提案された．本研究では，移動軌跡から単語の領域表現を教師なし学習によって獲得するために Word2Box を利用する．

3 移動軌跡への Word2Box の適用

3.1 移動軌跡の変換

移動軌跡を Word2Box で学習可能な形式に変換する方法について説明する．移動軌跡 $T = \{p_1, p_2, \cdots, p_n\}$ は，測位点 p_i のシーケンスで構成される．そして，測位点 $p_i = \{lat_i, lng_i, t_i\}$ は，GPS で測位された緯度 lat_i，経度 lng_i，およびタイムスタンプ t_i で構成される．

移動軌跡に対する前処理として，連続する 2 つの測位点のタイムスタンプの差が閾値より大きい場合は，その測位点の前後で移動軌跡を分割する．分割した移動軌跡について，一定回数の移動をしていない移動軌跡を削除する．また，測位点のフィルタリングとして，連続する 2 つの測位点の移動速度が閾値よりも大きい場合はその測位点を削除する．

本研究では，地域メッシュを表現するために quadkey[2] を用いる．quadkey は，地図上の特定の地域を一意に識別するためのインデックスある．特定のズームレベルによって作成されたインデックスの値は，重複する地域が存在しない四角形の地域メッシュとなる．なお，本研究では quadkey を利用しているが，地域の重複がない地域メッシュに変換可能であれば，quadkey 以外の手法や，四角形以外の形状でも構わない．移動軌跡 T のそれぞれの測位点を quadkey の値のシーケンスに変換した結果を $T' = \{w_1, w_2, \cdots, w_m\}$ とする．w_i は，測位点 p_i から特定のズームレベルで作成された quadkey の値であ

1） https://twitter.com/

2） https://learn.microsoft.com/azure/azure-maps/zoom-levels-and-tile-grid

表 1: データセット

投稿期間	2017/01/01 ~ 2018/12/31
ポスト数	3,366,896
ユーザ数	49,179
移動軌跡数	148,846
メッシュ数	3,301,496
メッシュの種類数	26,509

る．このとき，同じインデックスが連続することがあるが，それは地域メッシュ間の移動ではないとみなし重複を削除する．この処理により得られた quadkey の値を単語，移動軌跡を文とみなして，Word2Box を適用する．

3.2 Word2Box

本研究では，Word2Box を適用することで，Box Embeddings による領域表現を獲得する．はじめに，Box Embeddings について説明する．Box Embeddings は，d 次元空間において，単語 w の領域を 2 つの点で表す．単語 w の Box Embeddings による領域 $Box(w)$ を以下のように表現する．

$$Box(w) := \prod_{i=1}^{d} [x_i^-, x_i^+] = [x_1^-, x_1^+] \times \cdots \times [x_d^-, x_d^+] \quad (1)$$

ここで，x_i^- と x_i^+ は，それぞれボックスの左下と右上の座標を表し，d 次元のベクトルである．

次に，Word2Box の学習について説明する．Word2Box は，CBOW のように学習し，単語の領域間のマージンの最適化に取り組む．はじめに，文の単語から中心語を 1 つ選択する．また，ウインドウサイズに従い周辺語の集合を作成する．そして，それぞれの単語を Box Embeddings に変換し，中心語と周辺語の領域の重なりが大きくなるように学習を行う．また，ネガティブサンプリングによる負例を作成し，中心語と負例の領域の重なりが小さくなるように学習を行う．この学習結果に対して，2 つの単語の領域の共通部分や差を取ることで，集合演算を行うことが可能である．

4 地域メッシュの集合演算

4.1 データセット

本研究では，X から得たポストをデータセットとする．データセットが含む全てのポストは，東京都周辺で投稿された緯度経度情報が付与されている．データセットの詳細を表 1 に示す．

本研究では，quadkey のズームレベルは 17 と設定した．このズームレベルにより作成される地域メッシュの一辺は，約 305m である．Word2Box のパラメータについて，Box Embeddings に用いる 2 つの点の次元数は，それぞれ 32 次元とした．また，ウインドウサイズは 20 と設定した．演算結果の地域メッシュの可視化には，OpenStreetMap[3] を用いた．

4.2 集合演算の結果の可視化

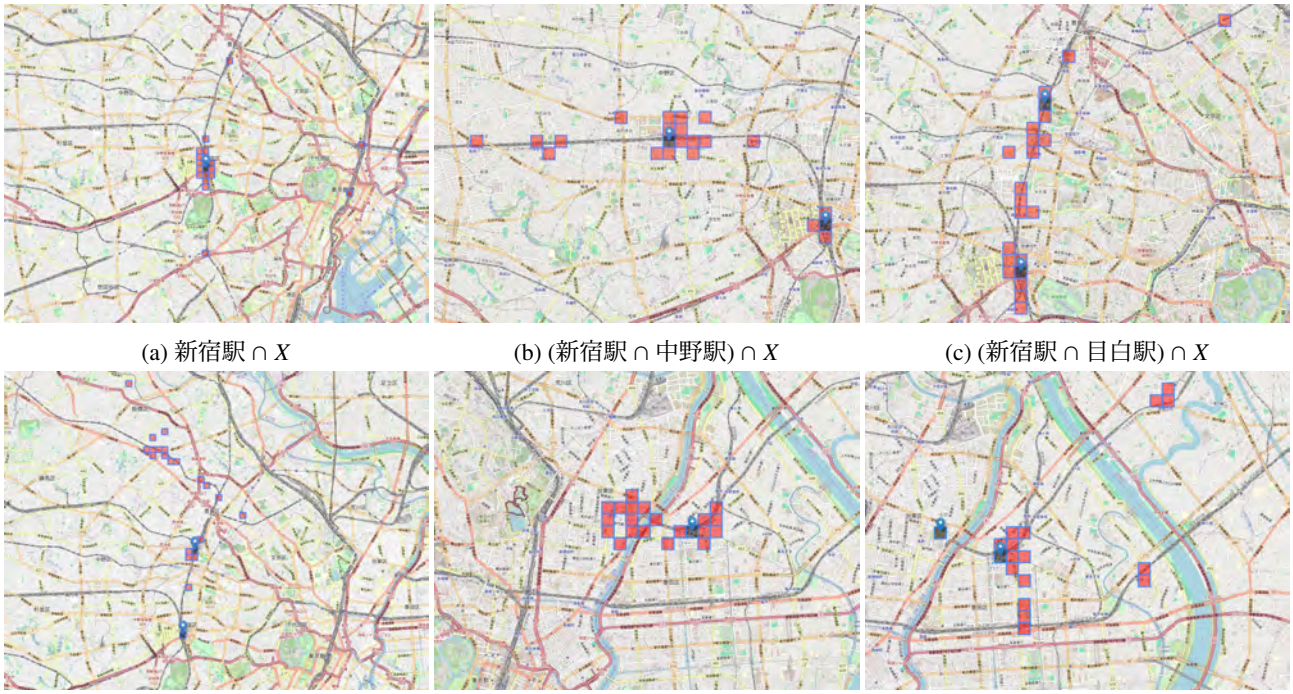
Word2Box を用いて獲得した領域表現の集合演算を行った結果について示す．集合演算により得られた領域の大きさが集合演算のスコアとなる[4]．たとえば，図 2 (a) の 新宿駅 ∩ X は，新宿駅を含む地域メッシュとそれ以外の地域メッシュ X の積集合のスコアを計算することを意味する．また，図 2 (d) の (目白駅 \ 新宿) ∩ X は，目白駅と新宿駅のそれぞれを含む地域メッシュの差集合を計算した結果とそれら以外の地域メッシュとの積集合のスコアを計算することを意味する．

図 2 (a) ～ 図 2 (f) に，集合演算により得られた地域メッシュを示す．それぞれの図において，マーカのある黒い地域メッシュが演算に用いた地域メッシュである．また，赤いメッシュが演算結果で得られたスコアが高い上位の 20 件である．

図 2 (a) において，スコアが高い地域メッシュとして，新宿駅の周辺や新大久保駅が得られた．この結果は，それらの地域メッシュの往来は，新宿駅に類似していることを示している．また，東京駅や，池袋駅などの主要な駅の地域メッシュも得られている．これらの駅は，多様な地域メッシュとの往来があることからそれぞれの領域が大きく，それらの地域メッシュの領域の大きさは，すべての地域メッシュの中で東京駅が 1 位，新宿駅が 3 位，および池袋駅が 83 位である．結果として，それらの共通領域が大きくなったと考えられる．

図 2 (b) と図 2 (c) は，新宿駅の地域メッシュと，中野駅と目白駅の積集合をそれぞれ求めた領域のスコアが高い地域メッシュである．中野駅と目白駅のいずれの駅も新宿駅からの鉄道路線がある駅である．この集合演算は，新宿駅との往来と，それらの駅との往来に共通する往来を持つ地域メッシュを得ることを表す．図 2 (a) と比較して，図 2 (b) と

3） https://www.openstreetmap.org
4） 計算方法の詳細は，論文 [11] を参照されたい．



(a) 新宿駅 ∩ X　(b) (新宿駅 ∩ 中野駅) ∩ X　(c) (新宿駅 ∩ 目白駅) ∩ X

(d) (目白駅 \ 新宿駅) ∩ X　(e) 東京スカイツリー ∩ X　(f) (東京スカイツリー \ 浅草寺) ∩ X

図 2: 集合演算のスコアが高い地域メッシュの上位の 20 件

図 2 (c) は，それらの駅が存在する中央本線の駅や，山手線の駅とそれらの周辺の地域メッシュが得られており，新宿駅とそれぞれの駅の往来に類似する地域メッシュを得られたと考えられる．

さらに，図 2 (d) は，目白駅の地域メッシュから新宿駅の地域メッシュの差集合の結果との積集合のスコアが高い地域メッシュである．この集合演算は，目白駅との往来とは類似しているが新宿駅との往来とは類似していない地域メッシュを得ることを表している．この結果では，新宿駅周辺の地域メッシュは得られずに，東武東上線の駅とその周辺の地域メッシュが多く得られている．

次に，観光スポットの地域メッシュの集合演算を行った結果について述べる．図 2 (e) は，東京スカイツリーを含む地域メッシュと往来が類似する地域メッシュである．近隣に浅草寺や，とうきょうスカイツリー駅，押上駅などが存在するため，それらの周辺の地域メッシュが類似する地域メッシュとして得られている．図 2 (f) は，図 2 (e) の集合演算に浅草寺の本堂を含む地域メッシュの差集合の演算を加えた結果である．演算結果は，浅草寺周辺の地域メッシュは得られずに，錦糸公園や，複合商業施設であるオリナス，東京メトロ半蔵門線の錦糸町駅などを含む地域メッシュが得られている．そのため，浅草寺と組み合わせて往来されることが少なく，とうきょうスカイツリー駅などと共通した往来があると思われる地域メッシュを得ることができた．これらの結果より，移動軌跡から地域メッシュの領域表現を獲得することで，人々の移動に基づく地域メッシュの集合演算を行うことができたと考えられる．

5　まとめ

本研究では，移動軌跡から作成した地域メッシュの単語列に対して Word2Box を適用することで地域メッシュの領域表現を作成した．獲得した領域表現を用いて集合演算を行った結果を地図上で可視化し，定性的な分析を行った．

今後の課題として，今回作成した領域表現を地域メッシュの埋め込み表現を用いる他のタスクにおいて活用することがあげられる．また，今回は，Word2Box の学習の際に，地域メッシュを単語として学習しているが，quadkey は地域メッシュの表現に階層性があるため，FastText などのようにサブワードを利用することで階層性を考慮した領域表現を作成することがあげられる．

謝辞

本研究は JSPS 科研費 19K20418 の助成を受けたものです．また，データセットをご提供いただいた 筑波大学 吉田光男 准教授に心より感謝いたします．

参考文献

[1] Shuang Wang, Bowei Wang, Shuai Yao, Jiangqin Qu, and Yuezheng Pan. Location prediction with personalized federated learning. **Soft Computing**, 2022.

[2] Y. Chen, H. Zhang, W. Sun, and B. Zheng. Rntrajrec: Road network enhanced trajectory recovery with spatial-temporal transformer. In **2023 IEEE 39th International Conference on Data Engineering**, pp. 829–842, 2023.

[3] Jie Hu, Shijie Cai, Tengfei Huang, Xiongzhen Qin, Zhangbin Gao, Liming Chen, and Yufeng Du. Vehicle travel destination prediction method based on multi-source data. **Automotive Innovation**, Vol. 4, No. 3, pp. 315–327, 2021.

[4] Yan Luo, Haoyi Duan, Ye Liu, and Fu-Lai Chung. Timestamps as prompts for geography-aware location recommendation. In **Proceedings of the 32nd ACM International Conference on Information and Knowledge Management**, p. 1697–1706, 2023.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In **Advances in Neural Information Processing Systems**, Vol. 26, 2013.

[6] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing**, pp. 1532–1543, 2014.

[7] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. **Transactions of the Association for Computational Linguistics**, Vol. 5, pp. 135–146, 2017.

[8] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, pp. 2227–2237, 2018.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, 2019.

[10] Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. Probabilistic embedding of knowledge graphs with box lattice measures. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 263–272, 2018.

[11] Shib Dasgupta, Michael Boratko, Siddhartha Mishra, Shriya Atmakuri, Dhruvesh Patel, Xiang Li, and Andrew McCallum. Word2Box: Capturing set-theoretic semantics of words using box embeddings. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2263–2276, 2022.

[12] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In **Proceedings of the 31st International Conference on Machine Learning**, Vol. 32, pp. 1188–1196, 2014.

[13] Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. emoji2vec: Learning emoji representations from their description. In **Proceedings of the Fourth International Workshop on Natural Language Processing for Social Media**, pp. 48–54, Austin, TX, USA, 2016.

[14] Aditya Grover and Jure Leskovec. Node2vec: Scalable feature learning for networks. In **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, p. 855–864, 2016.

[15] Jingcheng Du, Peilin Jia, YuLin Dai, Cui Tao, Zhongming Zhao, and Degui Zhi. Gene2vec: distributed representation of genes based on co-expression. Vol. 20, p. 82, 2019.

[16] Masaharu Hirota, Tetsuya Oda, Masaki Endo, and Hiroshi Ishikawa. Generating distributed representation of user movement for extracting detour spots. In **Proceedings of the 11th International Conference on Management of Digital EcoSystems**, 2020.

[17] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In **Advances in Neural Information Processing Systems**, Vol. 30, 2017.

[18] Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In **Proceedings of the 35th International Conference on Machine Learning**, Vol. 80, pp. 3779–3788, 2018.

[19] Alexandru Tifrea, Gary Becigneul, and Octavian-Eugen Ganea. Poincare glove: Hyperbolic word embeddings. In **International Conference on Learning Representations**, 2019.

[20] Luke Vilnis and Andrew McCallum. Word representations via gaussian embedding. In **3rd International Conference on Learning Representations**, 2015.

[21] Hongyu Ren and Jure Leskovec. Beta embeddings for multi-hop logical reasoning in knowledge graphs. In **Proceedings of the 34th International Conference on Neural Information Processing Systems**, 2020.

[22] Octavian Ganea, Gary Becigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In **Proceedings of the 35th International Conference on Machine Learning**, Vol. 80, pp. 1646–1655, 2018.

[23] Ryota Suzuki, Ryusuke Takahama, and Shun Onoda. Hyperbolic disk embeddings for directed acyclic graphs. In **Proceedings of the 36th International Conference on Machine Learning**, Vol. 97, pp. 6066–6075, 2019.

[24] Shib Dasgupta, Michael Boratko, Dongxu Zhang, Luke Vilnis, Xiang Li, and Andrew McCallum. Improving local identifiability in probabilistic box embeddings. In **Advances in Neural Information Processing Systems**, Vol. 33, pp. 182–192, 2020.