自然言語処理を用いた都市・建築の文字表現の類似性
Similality of Character Expressions in Citys and Architecture Using Natural Language Processing

○谷川 奈央[*1]，山田 悟史[*2]
Nao TANIKAWA[*1], Satoshi YAMADA[*2]

[*1] 立命館大学　理工学研究科　建築都市デザイン学科
Graduate School, Department. of Architecture and Urban Design, Ritsumeikan University.
[*2] 立命館大学　理工学部　建築都市デザイン学科　准教授・博士（工学）
Associate Professor. of Architecture and Urban Design, Ritsumeikan University. Dr.Eng

キーワード：自然言語処理；空間表現；類似性
Keywords: natural language processing; Spatial representation; similarity.

1. はじめに

文学における空間は，文字の羅列のみで存在し現実には存在しない空間である。しかし，その空間は登場人物の心情を映し出していたり場面の転換を意味づけたりと重要な役割を果たしている。空間描写があるからこそ物語が奥深く広がり，読者がさらに文学世界に没頭するのである。それらは，文字で表現された優れた空間デザインである。また，既存の空間を描いた文学表現は，類まれな感性がとらえた表現対象の空間特性である。それゆえに，文学作品中における空間表現の読解は，形態表現として空間をデザインする際の優れた参照対象になりえる。実際に複数の既往研究がある。特に空間表現が卓越している文学者として夏目漱石が知られている。若山らの研究[1]では作中の舞台空間の遷移から作品類型を作成している。他にも若山は著書[2]において，漱石の作風変化は東京という都市の変化を反映している，空間描写が時代を反映していると述べている。他にも村上春樹作品における空間を図学的観点から考察した研究として浜田の研究[3]がある。しかし，これらの研究は研究者自身が大量の作品を読み，該当部分を抽出し，整理，読解することで行われている。これは言語データを機械で処理することが困難だったためである。しかし，近年では，自然言語処理技術の発達により大量のテキストデータを高速に処理することが可能になった。例えば，Word2Vec[4],fastText[5],BERT[6]などである。Word2Vec は膨大なテキストデータを学習データとして単語を分散表現を獲得する手法である。fastText も同様の手法であるが，Word2Vec の対象が単語であったのに対し，fastText は単語の内部構造である語幹を元に分析する手法である。これにより人間の感覚に近い認識で単語の分散表現を獲得することが可能になった。BERT は転移学習の要素を導入したニューラルネットワークモデルである。通常ファインチューニングをすることで利用され，さらに高精度な処理が可能になった。単語の分散表現を獲得するという意味では Word2Vec や fastText と同様だが，大きな違いは BERT では文脈依存であるという点である。つまり，BERT では同じ単語であっても別の分散表現を獲得することが可能になる。

これらの自然言語処理技術を利用し，先人たちが残したテキストを読解することで新たな価値ある知見が生まれる可能性がある。本研究は，司馬遼太郎著書の『街道をゆく』[7]を対象に自然言語処理技術を用いてそれぞれの街道の類似性を把握し，街道ごとの特徴や空間表現の共通性や相違性を考察することを目的とする。

2. 研究概要

『街道をゆく』は著者が諸街道を旅しながら独自の視点からその地域の歴史・地理・人物について考察している紀行文集である。既往研究として山崎らの研究[8]がある。この研究では文章構成に着目することで表現手法ををを分析している。本研究では，Python の自然言語処理ライブラリである gensim の Word2Vec,Doc2Vec[9]を使用する。テキストデータに対しはじめに Doc2Vec を用いて他のテキストと特に類似しているテキストはどのテキストであるのか，どのテキスト同士が類似しているのかを把握する。その後，Word2Vec を用いて単語の類似度を測ることで各街道の共通点・相違点を考察していく。

2.1 データセット

『街道をゆく』全 43 巻のうち日本国内で対象地域が限定されているものを研究対象とした。各巻はいくつかの章に別れているため，一つの章を一つのテキストデータとした。データセット一覧が表 1，各テキストのおおよその該当エリアを地図上に示したしたものが図 1 である。

2.2 使用モデルと学習条件

Word2Vec は前述の通り単語の分散表現を獲得する手法である。Doc2Vec は同様の手法を文章に用いた手法であり文章 ID を付与することにより文章の分散表現を獲得す

ることが可能である。

各モデルの設定は以下の通りである。

・Word2Vec

出力されるベクトルの次元を 100, ウィンドウ幅[注1)] を 5 に設定 ,skipgram[注2)] で学習を行い学習回数は 100 回に設定。テキストごとに学習を行いそれぞれモデルを作成。

・Doc2Vec

出力されるベクトルの次元を 100, ウィンドウ幅を 7 に設定，dmpv[注3)] で学習を行い学習回数を 100 回に設定 , 全てのテキストデータを学習しモデルを作成。

3. 結果と考察

3.1 結果

該当テキストを全て学習させ作成した Doc2Vec モデルでテキスト間の類似度を測った。各テキストの類似度の合計値を表したグラフが図 2, 各テキストの上位 5 つと下位 5 つと類似度合いを表したグラフが図 3 である。 棒グラフ上にはテキストナンバーが付与されている。

3.2 考察

図 2 より類似度の合計値が一番低い値を示したテキストは 16: 甲賀・伊賀のみちであることがわかる。このことは，他の街道とは違い特徴的な描かれ方がされているのではないかと推測される。16 について Word2Vec によりテキスト中の頻出単語上位 3 単語の上位の類似単語一覧が表 2 である。伊賀 / 甲賀 / 峠の順である。伊賀・甲賀は，日本随一の歴史的な忍者の町であるとされ他の街道と違った要素を持ち合わせていることはほぼ間違いないと考えられる。頻出単語のうち伊賀・甲賀は類似度が一番高い " 衆 " と合わせて忍者集団の俗称である伊賀衆・甲賀衆を指す。" 峠 " は類似度が一番高い単語が " 斎 " であることから伊賀と甲賀をつなぐ「御斎峠」を指していると考えられる。この峠は著者の長編小説である『梟の城』の最初の場面の場所である。作品中に "「梟の城」を執筆する際に伊賀盆地を訪れ御斎峠に出かけたが，ひどく遠く，その上道が悪くて雨の日に登れるような場所ではないとあきらめ，結局遠望するのみで，地図を読みつつ書くことにした。" と記述されている。つまり，今回の伊賀訪問は筆者にとって念願の御斎峠であったと考えられる。これらを加味すると御斎峠は著者にとって特別な想いがある場所であると捉えることができる。頻出単語の類似語の中身をみてみると，伊賀・甲賀は半数ほどが一致しており，互いの類似語にお互いが存在していることから近い文脈で使われていることがわかる。

" 鈎 "," 城 "," 国 "," 戦国 " などの類似語が散見されることから街道の空間描写よりも伊賀衆・甲賀衆の歴史が語られていると推測できる。甲賀の類似語の中に " 高頼 "," 六角 " があるがこれは六角高頼のことである。高頼は室町幕府 9 代目将軍である足利義尚に攻め入られ甲賀の山中に逃げた際に甲賀衆たちに匿われた人物である。最終的

Table 1: List of Research Text

No.	タイトル	該当地域	巻数
0	湖西のみち	滋賀県	1
1	竹内街道	奈良県	
2	甲州街道	東京都	
3	葛城のみち	奈良県	
4	長州路	山口県	
5	睦のみち	青森県・岩手県	3
6	肥薩のみち	熊本県・鹿児島県	
7	河内みち	大阪府	
8	洛北街道	京都府北部	4
9	郡上・白川街道と越中諸道	岐阜県・富山県	
10	丹波篠山街道	京都府・兵庫県	
11	堺・紀州街道	大阪府	
12	北国街道とその脇街道	滋賀県・福井県	
13	那覇・糸満	沖縄県	6
14	石垣・竹富島		
15	与那国島		
16	甲賀・伊賀のみち	滋賀県・三重県	7
17	大和・壷坂のみち	奈良県	
18	明石海峡と淡路みち	兵庫県淡路島	
19	砂鉄のみち	四国北部	
20	熊野・古座街道	和歌山県	8
21	豊後・日田街道	大分県	
22	大和丹生川（西吉野）街道	奈良県	
23	種子島みち	鹿児島県	
24	潟のみち	新潟県	9
25	播州揖保川・室津みち	兵庫県	
26	高野山みち	和歌山県	
27	信州佐久平みち	長野県	
28	羽州街道	山形県	10
29	佐渡のみち	新潟県佐渡市	
30	蒙古塚・唐津	福岡県・佐賀県	11
31	平戸	長崎県	
32	横瀬・長崎		
33	五条・大塔村	奈良県	12
34	十津側		
35	壱岐・対馬の道	福岡県	13
36	函館・松前・江差の道	北海道	15
37	札幌・厚田・新十津川の道		
38	旭川・陸別の道		
39	叡山の諸道	滋賀県比叡山	16
40	芸備の道	広島県	21
41	神戸散歩	兵庫県	
42	横浜散歩	神奈川県	
43	近江散歩	滋賀県	24
44	奈良散歩	奈良県	
45	因幡・伯耆のみち	鳥取県	27
46	檮原街道（脱藩のみち）	高知県	
47	大徳寺散歩	京都府大徳寺	34
48	中津・宇佐のみち	大分県	

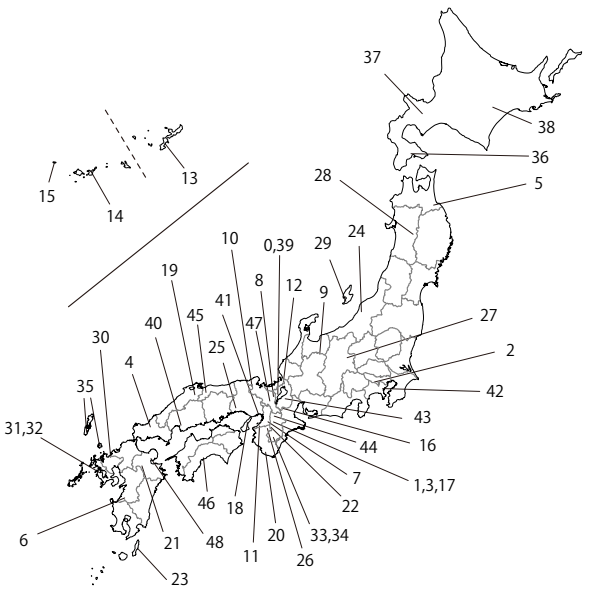

Figure1: Area corresponding to text number

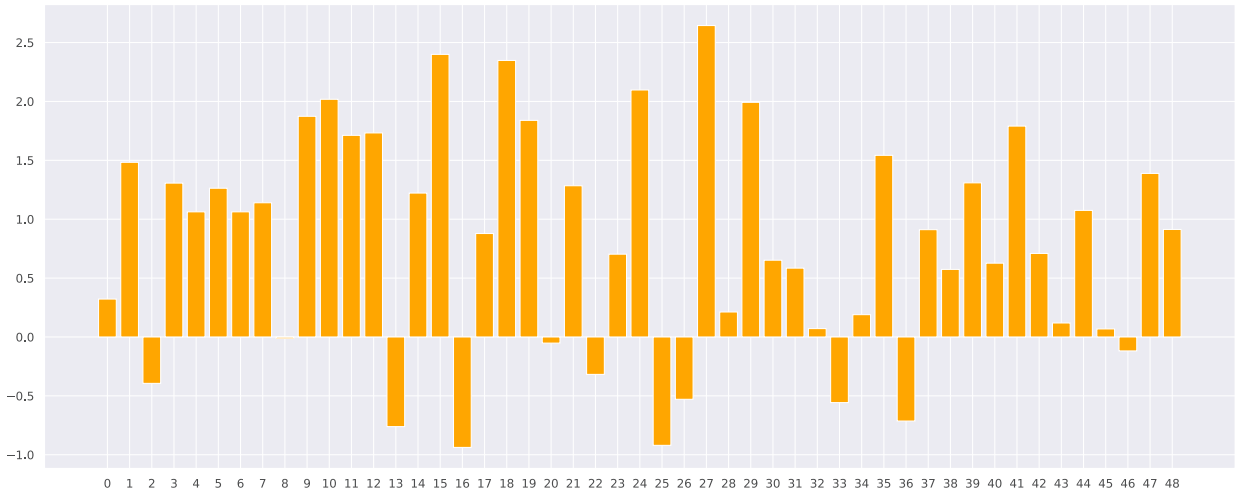

Figure 2: Total Value of Cosine Similarity for Each Text

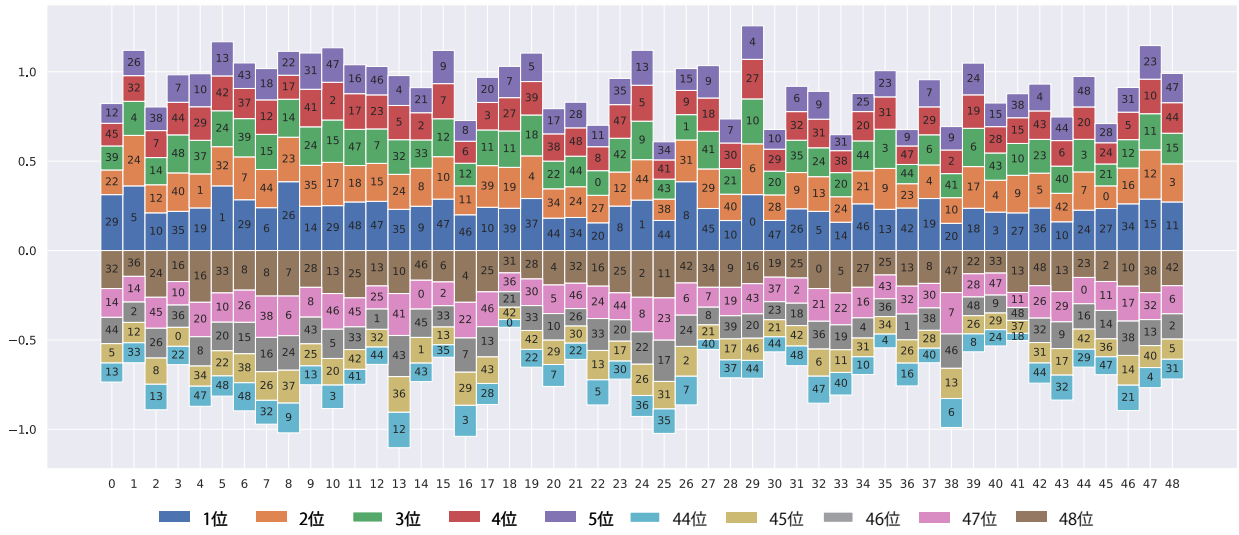

Figure 3: Doc2Vec Results Top5 and Bottom 5 for Each Text

Table 2: Similar Words of Frequent Word in 26 Text

伊賀		甲賀		峠	
衆	0.9477	衆	0.9524	斎	0.9917
多い	0.9472	戦国	0.9271	ゆく	0.9653
ほう	0.9381	近江	0.9097	よい	0.9603
たち	0.9230	多い	0.9090	いい	0.9583
国	0.9195	高頼	0.9042	のぼる	0.9393
者	0.9080	ころ	0.8993	道	0.9385
ころ	0.9041	出る	0.8985	老人	0.9308
郡	0.9003	期	0.8918	谷	0.9206
知る	0.8913	伊賀	0.8901	登る	0.9138
甲賀	0.8901	六角	0.8829	何	0.9109
出る	0.8779	義尚	0.8827	城	0.9054
城	0.8714	郡	0.8801	伊賀上野	0.9024
人物	0.8709	国	0.8715	あたり	0.9003
場合	0.8620	ほう	0.8660	入る	0.8949
鈎	0.8558	者	0.8627	盆地	0.8906
さ	0.8444	人物	0.8618	言う	0.8904
伊賀上野	0.8439	鈎	0.8518	どこ	0.8786
仕事	0.8376	思える	0.8491	石	0.8777
盆地	0.8372	たち	0.8385	中	0.8681
自分	0.8357	家	0.8362	自分	0.8667

Table 3: Top 20 Similar Words of " 街道 "

8		26	
風景	0.4322	口	0.6875
山国	0.4136	平安朝	0.6649
御陵	0.3813	慈尊院	0.5085
峠	0.3623	登る	0.5061
ほう	0.3405	俊	0.4886
世界	0.3395	よぶ	0.4869
つづく	0.3338	神社	0.4789
うち	0.3216	たどる	0.4698
背	0.3186	道	0.4691
登る	0.3163	離れる	0.4325
皇	0.3108	旧道	0.4255
陵	0.3088	途中	0.4193
よい	0.3067	乗	0.4172
鞍馬	0.3055	大門	0.4148
仲間	0.3038	うち	0.4145
軒	0.2996	高野	0.4078
杉	0.2968	東大寺	0.4024
釣れる	0.2933	土	0.4020
里	0.2874	百	0.3929
細流	0.2868	山頂	0.3887

に甲賀衆と高頼は栗田郡鈎に陣営を構えていた義尚に夜襲や奇襲を繰り返し勝利を収めた。(鈎の陣)これにより甲賀衆は世間に評判となったと言われている。甲賀の類似語の中に " 六角 "," 高頼 " が存在することはこの繋がりが記述されていたためであると推測される。このように，街道の空間描写ではなく，歴史的観点を主軸に語られていた，著者にとって思い入れのある場所であるという点から他のテキストとの差異が現れたのではないかと考えられる。

テキスト間で一番高い類似度を示した組み合わせは，8: 洛北街道と 26: 高野山のみちである。8 は京都市北部の鞍馬街道，周山街道の周辺を主な舞台としており街道の途中には山国陵がある。26 の高野山は空海が開いた真言宗の総本山金剛峯寺がある場所である。8 と 26 のテキストに対して，Word2Vec で " 街道 " に近いベクトルを持つ上位 20 単語を出力した結果が表 3 である。8 の類似語に " 峠 "26 の類似語に " 山頂 "," 道 "," 旧道 " という単語が上位に含まれておりさらに，どちらにも " 登る " という単語が存在している。このことからどちらの街道も山道でありその街道中に目的地がある共通性がうかがえる。しかし，8 の類似語に " 軒 "," 釣れる "," 里 "," 細流 " といった街道中の集落や自然を連想させるような単語が含まれるのに対して，26 の類似語では " 慈尊院 "," 神社 "," 大門 "," 東大寺 " といった寺社に関する単語が含まれており相違性がある。これは街道の性質の違いが類似単語に反映された結果だと考えられる。8 の街道中はわずかながら家々が軒を連ねており生活を営む人や宿の婦人と交流する様子が描かれている。さらに，街道中の自然の描写も散見される。一方 ,26 の高野山周辺の街道は総本山金剛峯寺へ向かうための街道であり街道中に人々の生活は営まれていない。テキスト中には街道沿いにある寺社の歴史や真言宗や平安仏教のその後の変遷が記述されている。

どちらのテキストも街道の地形に沿った空間描写がされている一方で，街道の性質の違いから " 街道 " の類似単語に違いが現れていることがわかる。

歴史的観点からみると，空海が高野山に開いた真言宗，最澄が比叡山に開いた天台宗は同じ平安仏教である。著者は 26: 高野山のみち ,39 叡山の諸道でそれぞれ総本山である高野山 , 比叡山とその周辺を訪れている。だが，同じ平安仏教という要素をもちながらも互いのテキストが類似しているという結果にはならなかった。むしろ図 3 から類似度が低いという結果になっていることがわかる。" 道 " の類似語をみると (表 4), 39 のテキストが最澄の " 澄 " や " 大宮 "," 大津 " といった人名や地名に関するものがある一方で，26 のテキストの類似語には地名などは含まれないことから道としての描かれ方に相違性があることがわかる。また， テキスト中に平安仏教に関する単語の出現回数を確認すると (表 5)，関連単語が出現しているが真言宗 / 真言 , 天台宗 / 天台など区別されていることがわかる。" 真言 " は " 真言密教 " や " 真言立川流 " などの単語が一単語として処理されず形態素分解された結果である。ほぼ同義にも関わらず別単語として認識されたことが類似度を低下させた一因とも考えられる。

Table 4: Similar words of " 道 "

26		39	
途中	0.7679	澄	0.4880
街道	0.6923	通用	0.4382
林	0.6784	大宮	0.4360
処	0.6404	橋	0.4321
入る	0.6381	登る	0.4285
とる	0.6258	川	0.4203
苅萱	0.6126	途中	0.4035
真	0.6102	くぐる	0.3923
別	0.6088	山口	0.3917
谷	0.6028	大津	0.3769

Table 5: Number of Occurrences of Words

26		39	
空海	47	最澄	139
高野山	44	比叡山	10
密教	32	法華経	18
真言宗	3	天台宗	44
真言	16	法華	44

Table 6: Top Similar Words

天台宗	
最澄	0.4548
真言宗	0.4458
大乗	0.4293

39 のテキストで作成した Word2Vec モデルで " 天台宗 " の上位の類似単語が表 6 より " 最澄 "," 真言宗 " であることがわかる。平安仏教に関する単語同士が近しい分散表現を獲得するモデルを該当のテキストデータのみから作成することができたということがわかった。

4. まとめ

自然言語処理技術を用いて『街道をゆく』を解析し，著者独自の視点から記述された諸街道の単語の類似度を把握することで街道の性質や空間表現の仕方を考察することができた。今後は fastText,BERT いった他のモデルを適用を試みる他に，既存の辞書にユーザー辞書を登録しテキストに合わせた詳細な分析を可能にしていきたい。

[注釈]

注 1) 対象単語から前後何単語考慮するかを指定する

注 2)skipgram: 対象単語から周辺の単語を予測する手法

注 3)dmpv: 語順を考慮しながら，周辺の単語から対象の単語を予測する手法に文章 ID を付加した手法

[参考文献]

1) 若山滋 , 張奕文 , 渡辺孝一 , 夏目漱石作品の中の建築の研究 - 舞台空間の推移からみた作品の類型について -, 日本建築学会計画系論文集 , 第 476 号 ,pp.101-109,1995

2) 若山滋 , 漱石まちをゆく - 建築家になろうとした作家 -, 彰国社 ,2002

3) 浜田真理 , 村上春樹の小説にあらわされた空間について - 図学的観点からの考察 -,2005 年度大会 (関東) 学術講演論文集 ,pp.125-128,2005

4)Tomas Mikoluv et al. Efficient Estimation of Word Representations in Vector space, ICLR Workshop,2013

5)Bojanowski Piorr,Edoward Grave,Armand Joulin, Tomas Mikoluv, Enriching Word Vectors with Subword,2016

6)Jacob D, Ming-Wei C, Kenton L, et al. ,BERT:Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the NAACL-HLT, 2019

7) 司馬遼太郎 , 街道をゆく (全 43 巻), 朝日文庫

8) 山崎隆之 , 十代田朗 , 地域イメージの表現手法に関する研究 - 司馬遼太郎の『街道をゆく』における文章構成の分析から -, 日本都市計画学会都市計画論文集 ,No.39-3,pp.97-102,2004

9) Quec Le, Tomas Mikoluv, Distributed Respresentetions of Sentences and Documents, Proceedings of the 31st International Conference on Machine Learning Vol.32, No.2,2014