修 士 論 文 概 要 書

Master's Thesis Summary

Date of submission: 01/10/2024 (MM/DD/YYYY)

専攻名（専門分野） Department	経営システム 工学専攻	氏 名 Name	森本　貫太 Kanta Morimoto	指 導 教 員 Advisor	後藤　正幸　　印 Seal
研究指導名 Research guidance	情報数理応用研究	学籍番号 Student ID number	5222C035-1		
研究題目 Title	宿泊施設を対象とした BERT と自動翻訳に基づく多言語レビューの埋め込み表現モデルに関する研究 Embedding based model for multilingual reviews based on BERT and automatic translation for accommodations				

1. はじめに

近年，日本に訪問する外国人旅行者数は増加傾向にあり，それに伴い，外国人旅行者向けの様々なサービス品質向上が求められている．そこで本研究では，宿泊施設を対象とした外国人旅行者向けのサービス品質向上について着目する．これまでも，宿泊施設のサービス品質向上に結びつく施策立案のため，宿泊予約サイトにおける施設のレビュー文やユーザ評価値を利用してさまざまな研究が行われてきた．しかし，日本語レビュー，及び英語レビュー (以下，日英レビュー) を同時に扱った研究は未だ数が少ない．日本の旅行者だけではなく，外国人の旅行者にも対応する場合，多言語のレビュー文を統一的に分析する必要がある．なぜなら， 同じ宿泊施設を評価する場合であっても， 日本人観光客の評価の観点と外国人の評価の観点は異なる可能性があるためである．例えば， 日本人観光客や外国人観光客のどちらかにターゲットを特化した戦略だけでなく， 大きなコストアップを避けつつ，日本人観光客の満足度を満たしながら， 外国人観光客へのサービス品質向上を満たす方策の検討をするためには， 日本人観光客と外国人観光客の観点の共通部分と差異の正しい認識が必要であろう．このような観点から， 日本語レビューと英語レビューを同じドメインで分析し， 両者の共通項と差異を分析可能なモデルは有用である．

そこで本研究では，日本語と英語で投稿されたレビュー文を統一的に分析するため，日英レビューを同一意味空間上に埋め込み分析する手法を提案する．具体的には，宿泊予約サイトにおける日英レビューを機械翻訳機を援用して，汎用自然言語モデルである BERT(Bidirectional Encoder Representations from Transformers)[1]， 及び次元削減法の一つである MDS(Multidimensional Scaling)[2] を利用することによって同一意味空間上に表現し，宿泊予約サイトにおける日英レビューの特徴の差異を分析する．その結果，宿泊施設側としては， 日本人観光客との評価の差異を正しく理解することで， 両者の特性を踏まえたサービス改善の施策を検討することが可能となる．

2. 準備

2.1. BERT

BERT とは Google が 2018 年に公開した言語モデルであり，Transformer の構造を組み込み，大規模なテキストデータによるモデルの事前学習によって高い汎用性と文脈理解によって注目されたモデルである．学習済みのモデルを利用することで，新たな入力文に対して埋め込み表現列を獲得することができる．ここで，埋め込み表現とは各文書を実ベクトルで表現し，意味が類似している文書は近く，類似していない文書は遠くなるように各文書がベクトルで表現される．そして， Devlin ら [3] は単一言語のみ扱う BERT を複数言語に同時に対応学習できる Multilingual BERT モデルへと拡張した．これは日本語，及びその他 109 言語の文章における埋め込み表現を同一意味空間上で獲得することが可能なモデルであり，言語ごとに異なるモデルを用意する必要がなく，多言語を埋め込み表現を一つのモデルで扱うことが可能である．一方で，日本語や中国語のような形態素分解の際に係り受け解析が必要な言語において，Multilingual BERT の精度はあまり高くないことが知られている．

2.2. NMDS

NMDS(Non-Multidimensional Scaling) とは，対象 i と対象 $j(i \neq j)$ の親近性 s_{ij} が順序尺度のデータとして入力された時に， 対象間の類似度を再現するようにユークリッド空間に全対象を布置し，出力する潜在表現手法である．すなわち，獲得したい潜在空間の次元を P として，対象 i の座標を $X_i = (x_{i1}, x_{i2}, ..., x_{iP})$，対象 j の座標を $X_j = (x_{j1}, x_{j2}, ..., x_{jP})$ とした時，対象の全ての組み合わせの類似度から， ユークリッド空間内に， 類似度の大きいペアはなるべく近くなるように， 類似度の小さいペアはなるべく離れるように各座標 X_i，X_j を求める手法である．具体的には， 求めるべき座標から計算される類似度を d_{ij}，対象 i と対象 j の真の類似度を d^*_{ij} として， d_{ij} と d^*_{ij} の差の 2 乗和が小さくなるように座標を求める．ここで， d_{ij} は式 (1) に示すミンコフスキー距離によって定義され，t はミンコフスキー定数と呼ばれる．式 (1) は $t = 1$

の時にマンハッタン距離，$t = 2$ の時にユークリッド距離を表す．

$$d_{ij} = (\Sigma_{m=1}^{P}|x_{im} - x_{jm}|^t)^{1/t} \quad (1)$$

この時，対象の空間布置が類似度を説明する度合いを表す評価基準として，式 (2) に示す Kruskal によって定義された STRESS 値と呼ばれる値が小さくなるように最適な座標を探索す．

$$STRESS = \sqrt{\frac{\sum\sum_{i<j}(d_{ij} - d_{ij}^*)^2}{\sum\sum d_{ij}^2}} \quad (2)$$

最適な座標探索の際は，ミンコフスキー定数 t を変化させながら，STRESS 値によって距離データの再現の度合いを評価する．

3. 提案手法

言語の異なるテキストデータを同一意味空間上で扱う手法である Multilingual BERT を使用することで，宿泊予約サイトにおける日英レビューを同時に扱うことは可能である．しかし，Multilingual BERT における日本語文書の分析精度は日本語で学習した BERT モデル (以下，日本語 BERT モデル) と比較して一般的に劣るということが石原ら [4] の研究により確認されている．これは，モデルの構造において日本語 BERT モデルは形態素解析器により単語分割を行って学習を行っているのに対して，Multilingual BERT における日本語文書の学習では，一文字一文字を一単語として分割してしまっていることが原因と考えられている．ここで，日本語文書の埋め込み表現に対して高い精度で且つ日英レビューを同一意味空間上に表現するために，英語レビューについては機械翻訳器により日本語に翻訳してから，日本語 BERT モデルを使用することで，獲得した意味空間から日英レビューの特徴の違いを分析することができると考えられる．しかし，この方法は，機械翻訳器を用いることで翻訳誤差のノイズが加わってしまうといった問題点がある．しかし，大規模なビッグデータを扱う場合はこのようなノイズを人手で処理するのは困難であるため，機械翻訳器による言語間のノイズを考慮した分析手法が必要になる．

そこで本研究では，多言語のレビュー文を統一的に分析するため，機械翻訳器による言語間のノイズを考慮して，日英レビューを同一意味空間上にて分析する手法を提案する．具体的には，汎用自然言語処理モデルである BERT，及び機械翻訳器を利用して，入力を全て日本語のテキストデータ (英語レビューを全て日本語に翻訳する) とする BERT 日本語モデル，入力を全て英語のテキストデータ (日本語レビューを全て英語に翻訳する) とする BERT 英語モデルの双方から日本語意味空間と英語意味空間を獲得する．次に非計量多次元尺度法である NMDS を利用して 2 つの言語における意味空間を組み合わせた 1 つの空間を再構成する．すなわち，提案手法は，異なる 2 つの意味空間を組み合わせて，機械翻訳器による言語間のノイズを考慮し，両モデルの整合性を持って新たな意味空間を構築する方法である．

提案手法における手順を以下の **Step1**∼**Step5** に示す．ただし，レビュー数を N として，t_i は i 番目のレビューデータ $(i = 1, 2, ..., N)$，$d_{t_i t_j}^{*'}$ は t_i と t_j の距離，$Jap\_sim(t_i, t_j)$ と $Eng\_sim(t_i, t_j)$ はそれぞれ BERT 日本語モデルと BERT 英語モデルにおける t_i と t_j のコサイン類似度，$ol(t_i)$ は t_i における元々の言語とする．

提案手法の分析ステップ

Step1 機械翻訳器を利用し，各レビューに対して日本語と英語の 2 種類のレビューを準備する．

Step2 事前学習済みモデルである日本語 BERT モデルと英語 BERT モデルにより，日本語と英語の 2 つの意味空間上に各レビューを埋め込む．

Step3 日本語意味空間上でのレビュー間の $N \times N$ 類似度行列 D_J^* と英語意味空間上でのレビュー間の $N \times N$ 類似度行列 D_E^* を生成する．ここで，類似度行列における各要素はレビュー t_i とレビュー t_j のコサイン類似度とする．

Step4 日本語と英語における 2 つの類似度行列 D_J^*, D_E^* から 1 つの平均類似度行列 $D^{*'} \in \mathbb{R}^{N \times N}$ を生成．平均類似度行列 $D^{*'}$ における要素 $d_{t_i t_j}^{*'}$ は式 (3) の条件式で与える．

Step5 平均類似度行列 $D^{*'}$ を入力とし，NMDS により，レビューにおける新たな 1 つの意味空間を獲得する．

Step4 において，条件式 (3) に示す通り，元の言語が違うレビューペアに対しては，その日本語レビューの類似度の英語レビューの類似度の平均を入力としている．こうすることにより，機械翻訳器におけるノイズの影響を和らげることが可能となる．また，元の言語が同じ文章ペアに対して日本語ペアと英語ペアの平均を入力としないのは，わざわざ両方とも翻訳を行い，2 つの類似度の平均を取ることで余計なノイズを加えることになるからである．

また，NMDS の入力における類似度は，本提案手法におけるレビュー間のコサイン類似度 (元の言語が異なる場合は英語と日本語におけるコサイン類似度の平均値) になる．本提案手法では，平均類似度行列 $D^{*'}$ を入力とする NMDS によって新たな 1 つの意味空間を生成することで，双方向の翻訳における異なる 2 つの意味空間において，言語間のノイズを考慮した，両モデルの整合性を持って新たな意味空間を構築することが可能となる．提案手法のイメージ図 1 に示す．

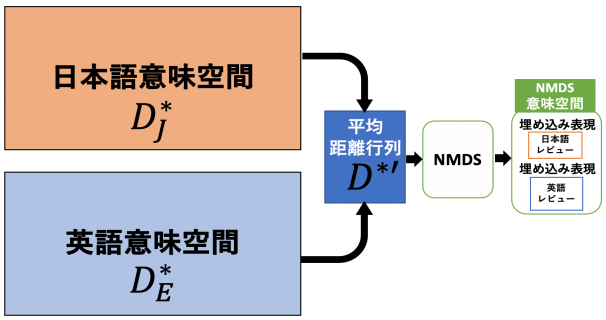

図 1. 提案手法のイメージ図

$$d^{*'}_{t_i t_j} = \begin{cases} Jap\_sim(t_i, t_j) \\ \quad if\ (ol(t_i) = Jap) \\ \quad\quad and \\ \quad (ol(t_j) = Jap) \\ \\ Eng\_sim(t_i, t_j) \\ \quad if\ (ol(t_i) = Eng) \\ \quad\quad and \\ \quad (ol(t_j) = Eng) \\ \\ \frac{1}{2}(Jap\_sim(t_i, t_j) + \\ Eng\_sim(t_i, t_j)) \\ \quad if\ ol(t_i) \neq ol(t_j) \end{cases} \tag{3}$$

4. 評価実験

4.1. 実験条件

本研究は，国立情報学研究所の情報学研究データリポジトリーで提供されている楽天トラベルデータセット (2019 年)[5] のユーザレビューを対象とする．データ数は日本語 83,188 件，英語 1,487 件であり，使用する日本語 BERT モデルは東北大学の乾・鈴木研究室の事前学習済み日本語 BERT モデル英語 BERT モデルは Bert-base-uncased モデル， 機械翻訳機には DeepLAPI を使用した．

4.2. 実験手順

本節では実データ実験により，提案手法の有効性を示す．提案手法の有効性は，レビュー文書間の類似度を正しく測ることができているか否かによって評価することが可能である．しかしながら，レビュー文書に対して“真の類似度”を知ることはできないため，評価実験には工夫が必要である．そこで，日本語レビュー間の類似度は日本語 BERT モデルで測った類似度が，英語レビューについては英語 BERT モデルで測った類似度が，本来，得たい類似度 (正解類似度) であると仮定して，自動翻訳器を使って言語を変換したレビューを混ぜたデータから，どの程度，この正解類似度を再現できるかを評価する．以下に実験の具体的な手順 1~ 手順 4 を示す．実験全体のイメージを図 2 に示す．

手順 1 実データからランダムサンプリングした日本語レビューデータ 1,000 件を仮の正解データとし，そのうちランダムに 500 件を選んで英語に翻訳してこれらを英語データ (仮) と仮定する．

手順 2 日本語データ 500 件， 英語データ (仮)500 件を使用して従来手法における BERT の日本語意味空間，及び英語意味空間，そして提案手法における意味空間の合計 3 つの意味空間を生成する． この時， NMDS における潜在表現の次元数は， BERT によって獲得する意味空間に合わせるため， 768 次元する．

手順 3 正解データの日本語 BERT モデルによる類似度行列と生成した 3 つの意味空間における類似度行列との各要素ごとの絶対誤差の平均，すなわち MAE(平均絶対誤差) を算出し， 提案手法がどれだけ正解データと近しい意味空間を再構築できているかを評価する．

手順 4 仮の正解データ 1,000 件を英語レビューとした場合について手順 1~3 を日英逆向きにして行う．

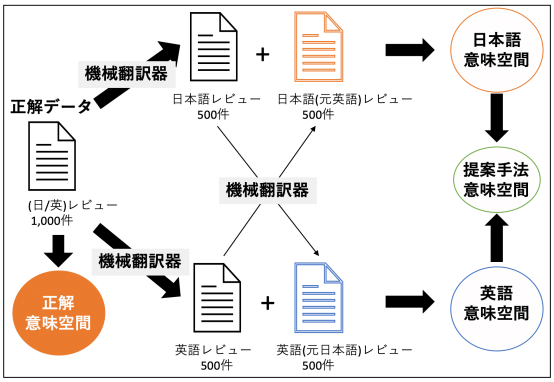

図 2. 実データ実験のイメージ図

4.3. 実験結果

日本語と英語それぞれにおける正解データの意味空間の類似度行列と生成した 3 つの意味空間 (提案手法，日本語 BERT モデル，英語 BERT モデル) における類似度行列との各要素ごとの絶対誤差の平均，すなわち MAE(平均絶対誤差) を比較した結果をそれぞれ表 1， 2 に示す．

表 1. 正解データが日本語の場合の MAE 比較

	日本語意味空間	英語意味空間	提案手法
MAE	1.87×10^{-6}	3.06×10^{-6}	2.66×10^{-6}

表 2. 正解データが英語の場合の MAE 比較

	日本語意味空間	英語意味空間	提案手法
MAE	2.97×10^{-6}	1.62×10^{-6}	2.49×10^{-6}

日本語レビューと英語レビュー， どちらを正解データとした場合も， 提案手法の意味空間は非正解データの言語における意味空間よりは誤差が小さく， 正解データの言語における意味空間よりは大きい結果となった．

5. 考察

5.1. 意味空間の再構成における精度について

日本語と英語の 2 つの意味空間の精度は，それぞれ正解である元の言語の意味空間よりも低いが，それらを組み合わせることで中和できていると言える．言い換えれば，本提案手法を利用することで，機械翻訳器による翻訳時に，日本語又は英語のどちらかの意味空間が持つノイズの影響を低減させることが可能であると期待される．

5.2. 提案手法意味空間によるレビュー分析

本節では，本提案手法によって再構成された意味空間を利用してレビュー内容を分析した結果について説明する

5.2.1. 朝食に関するレビュー

朝食に関するレビューの分布は日本語と英語共に類似しており， 共通した観点として， 宿泊施設を肯定的に評価する際は， 朝食の種類の多さに関することが多いことが分かった． 表 3 に抜粋したレビューの例を示す．

表 3. 朝食の品数に関するレビュー例

言語	レビュー
日本語	“朝食の品数も多く， 味も” “食べ物の品数が多く”
英語	“Variety menu of western and Japanese” “great varieties of breakfast”

一方， 英語ユーザーは日本語ユーザーに比べ， 朝食の料金や無料であることをより重視する傾向があることも分かった． 以下に英語レビューにおける抜粋した例を示す．

- we have no complaint because with this price the hotel provide free breakfast
- nice hotel with free breakfast

5.2.2. 英語対応に関するレビュー

また，宿泊施設に関する外国人旅行客へ向けた “ 英語 ” の取り扱いの有無について書かれたレビューの分布において，レビューの件数は日本語の方が多いものの，全体のレビュー件数に対する割合は英語レビューの方が多いことが分かった．ここで “ 英語 ” のに関する具体的なレビューの内容の例を表 4 に示す．日本人は, 同じ宿泊施設に滞在している外国人へ向けた配慮が足りない点に関するネガティブなコメントが多い一方で，外国人旅行者 (英語レビューワー) に関しては “ 英語 ” の取り扱いに関するネガティブな評価は少なく, むしろ英語を話せるスタッフを備えている宿泊施設を高評価している傾向にあることが窺える．

以上のことから， 宿泊施設側は以下の 2 点に注力をすることが，今後の外国人旅行者向けのサービス品質向上に向けた重要課題と考えられる．

表 4. 朝食の品数に関するレビュー例

言語	レビュー
日本語	“英語の注意事項が曖昧で湯船にタオルを” “外国人のために説明表示が図付きで”
英語	“is able to communicate in English” “staffs are nice with good English”

- 朝食に関して， より多い品数の量を保ちつつ， コストをできるだけ抑えた朝食提供
- スタッフの語学力強化や館内の英語表記の追加などのグローバル化の推進

6. 結論と今後の課題

本研究では， 宿泊予約サイトにおける日英レビューを同一意味空間上にて分析するために， BERT の日本語と英語の両モデルにより得られた 2 つの意味空間を NMDS によって再構成して 1 つの意味空間を構成する方法を提案した． 再構成した意味空間から日本語と英語のレビュー内容の比較分析を行った結果， 宿泊施設における海外旅行者向けのサービス品質改善のための重要な観点を明らかにした．

一方， 実データ実験の結果から，精度面では改善の余地がある．これは，NMDS の手法上，例え入力の類似度が限りなく近似している場合においても，得られる潜在表現が同一にはなり得ない性質を持っているためであると考えられる．そのため，より適切な方法で 2 つの意味空間を組み合わせる方法を考案することが今後の課題として挙げられる．

謝辞

本研究は，国立情報学研究所情報学研究データリポジトリにより楽天グループ株式会社から提供を受けた「楽天トラベルデータセット」[5] を利用した．ここに記して謝意を表す．

参考文献

[1] Kenton Lee Jacob Devlin, Ming-Wei Chang and Kristina Toutanova. Ert: Pre-training of deep bidirectional transform-ers for language understanding. *arXiv:1810.04805*, 2018.

[2] J. B Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, Vol. 29, No. 1, 1964.

[3] Dan Garrette Telmo Pires, Eva Schlinger. How multilingual is multilingual bert? *arXiv:1906.01502*, 2019.

[4] 石原慧人，石原祥太郎，白井穂乃． Bertsum を用いた日本語ニュース記事の抽象型要約手法の検討． *The 35th Annual Conference of the Japanese Society for Artificial Intelligence*, 2021.

[5] Informatics Research Data Repository National Institute of Informatics. (dataset) Rakuten Group, Inc. (2014). Rakuten Dataset. `https://doi.org/10.32130/idr.2.0`,.