卒業論文

異言語間単語埋め込みを用いた

日英間の文化差検出

指導教官　　村上 陽平 教授

立命館大学　情報理工学部
先端社会デザインコース　4 回生
2600200377-4

増田 憲都

2024 年度（秋学期）卒業研究 3（CH）
令和 6 年 1 月 31 日

異言語間単語埋め込みを用いた日英間の文化差検出

増田 憲都

内容梗概

近年，機械翻訳の品質が向上してきており，機械翻訳を用いた多言語コミュニケーションや異文化コラボレーションが可能になってきている．このようなコミュニケーションでは，翻訳は正しいものの，文化差によってその単語の表す概念に付随する属性やその概念に関連する概念が異なり，コミュニケーションの齟齬が生まれてしまう場合がある．例えば，日本では「鯨」食べる文化があるが，食用としない文化圏もある．鯨を食用とする文化圏の人と鯨を食用としない文化圏の人が話し合う時に，お互いに食習慣が尊重されていないと感じ，円滑にコミュニケーションが取れない場合がある．このようなコミュニケーションのずれを解消するために，画像検索を用いた対訳ペア間の文化差検出がある．しかしながら，この手法は画像検索の結果に依存するため，主要な画像しか検索されず，文化差のある画像が取得できない場合に正しく機能しない．また，対象概念に関連する概念に文化差がある場合，画像に大きな違いがないため，文化差を検出できない．

そこで，画像には表れにくい文化差を検出するために，概念を単語の分散表現を用いて表し，異言語間で単語の分散表現を対応づけることで文化差を検出する手法を提案する．具体的には，言語ごとの Wikipedia コーパスを用いて各言語の分散表現空間を取得して，異言語間単語埋め込みを使用して両空間のアライメントを行い，対応が取れない対訳関係の言語ペアを文化差として抽出する．本手法の実現にあたり，取り組むべき課題は以下の 2 点である．

異文化間での概念の同一性

日英の単語の分散表現のアライメント後に，対訳関係にある単語が同一の概念かどうか判定するには，類似度の閾値を定める必要がある．基準となる対訳関係の単語ペアを選択し，その全ての単語間の類似度が閾値を超えるように設定する必要がある．

関連概念の分析

アライメントにより対訳関係にある単語ペアの分散表現を一致させることができないとしても，その単語の概念が異文化間で異なるのではなく，アライメントに失敗した可能性がある．そこで，単語に関連する概念間の比較を通して，関連概念間でも差異があるのか分析する必要がある．

一つ目の課題に対しては，まず日英の単語の分散表現のアライメントを行う．

具体的には，Word2Vec を用いて生成した日本語と英語それぞれの分散表現に対して，敵対的なトレーニングによりネットワークが各言語の埋め込みを区別できるようにトレーニングする．さらに Procrustes 改良を利用し各言語の埋め込みの間で最適な対応を見つけ，日本語表現空間を中心として英語分散表現空間を回転させる．次に，k-近傍マッチングを行い，パラメータ k の値を変化させて日英間の対訳ペアの一致率を測定する．一致率が 90%以上になった時に一致しなかった対訳ペアを文化差候補として抽出する．さらに，語彙統計学で用いられる基礎語彙（スワデシュリスト）を用いて，各単語の分散表現が一致するよう単語間の類似度の閾値を定め，閾値以下の単語ペアに文化差候補を絞り込む．スワデシュリストとは，言語間の近縁度を測定するために，借用語とならない幼年期に学習される基礎的な単語であり，他の概念との関連に言語間で大きな差がない単語と考えられ，アライメント後の概念の同一性の基準として用いる．

二つ目の課題に対しては，アライメント後に一致しなかった各単語ペアの関連概念を上位 10 件抽出し，各関連概念を表す単語が異言語間で一致しているかどうか分析を行う．本研究の貢献は以下の通りである．

異文化間での概念の同一性

k-近傍法により，k=5 から k=7 の間で未発見の単語かつ同義語の単語ペアは（豆 bean)・(ベビーベッド crib)・(菌類 fungus)・(亀 turtle)・(通気孔 vent）だった.文化差がない基準としてスワデシュリストの cos 類似度を測った．（人 human）の cos 類似度 0.444 が最も低かったので，この数値を閾値として扱う．この結果，提案手法により文化差として抽出された単語のペアは（イチジク fig），（ベビーベッド crib），（通気孔 vent）だった.

関連概念の分析

関連概念の範囲として単語の cos 類似度上位 10 個に設定した．k-近傍法で上位 10 件を取得した結果，（イチジク fig），（ベビーベッド crib），（通気孔 vent）は直接的にあるいは間接的に該当の単語に一致していた．しかし，英語分散表現では明らかに関係のない単語も散見された.

Detection cultural differences between Japanese and English using cross-language word embedding

Kazuto Masuda

Abstract

Recently, the quality of machine translation has been improving, making multilingual communication and cross-cultural collaboration using machine translation possible. In such communication, although the translation is correct, cultural differences can cause discrepancies in communication due to differences in the attributes associated with the concept represented by the word and the concepts related to that concept. For example, in Japan, there is a culture that eats "whale", but there are other cultures that do not eat whale. When people from whale-eating cultures and people from non-whale-eating cultures talk to each other, they may feel that they are not respecting each other's food habits and may not be able to communicate smoothly. In order to resolve such a communication gap, there is a method of detecting cultural differences between bilingual pairs using image retrieval.

In order to detect cultural differences that do not appear in images, we propose a method to detect cultural differences by representing concepts in terms of word variants and mapping word variants between different languages. Specifically, we obtain the distributed representation space of each language using the Wikipedia corpus for each language, align the two spaces using interlingual word embedding, and extract language pairs that have no correspondence as cultural differences. In order to realize this method, the following two issues need to be addressed.

Identity of Concepts Across Cultures

To determine whether or not words in a bilingual relationship are the same. concept after alignment of distributed representations of Japanese-English words, a similarity threshold needs to be defined. It is necessary to select a reference pair of bilingual words and set the threshold so that the similarity between all the words exceeds the threshold.

Analysis of Related Concepts

Even if alignment fails to match the distributed representation of a pair of words in a bilingual relationship, it is possible that the alignment failed,

rather than that the concepts of the words differ between different cultures. Therefore, it is necessary to analyze whether there are differences even among related concepts through comparison between concepts related to the word.

For the first task, we first align the distributed representations of Japanese and English words. Next, k-nearest neighbor matching is performed, and the agreement rate between Japanese-English bilingual pairs is measured by changing the value of the parameter k. When the agreement rate reaches 90% or more, unmatched bilingual pairs are extracted as candidate cultural differences. Furthermore, using a basic vocabulary used in lexicostatistics (Swadeshi list), a threshold of similarity between words is set so that the distributed expression of each word matches, and candidate cultural differences are narrowed down to word pairs that are below the threshold.

For the second task, we extract the top 10 related concepts for each word pair that is not matched after alignment, and analyze whether the words representing each related concept are matched across different languages. As in the first task, the k-nearest neighbor method and word-to-word similarity are used to determine agreement. The contributions of this study are as follows.

Identity of Concepts Across Cultures

The words that are candidates for cultural differences, i.e., words that are undiscovered at k=5 to k=7 by the k-nearest neighbor method, are listed in Table 1. We measured the cos-similarity of the Swadeshi list as a criterion for the absence of cultural differences. (human) had the lowest cos similarity of 0.444, so we treat this value as the threshold value.

This value is treated as the threshold value. The word pairs that are lower than the threshold are (fig fig), (crib crib), (vent vent).

Analysis of Related Concepts

We set the top 10 words in terms of cos-similarity as the range of related concepts. The results of the k-nearest neighbor method showed that (fig fig), (crib crib), and (vent vent) were directly or indirectly related to the corresponding words. However, there were some words that were clearly unrelated in the English distributed expressions.

目次

第1章 はじめに

近年，機械翻訳の品質が向上してきており，機械翻訳を用いた多言語コミュニケーションや異文化コラボレーションが可能になってきている．このようなコミュニケーションでは，翻訳は正しいものの，文化差によってその単語の表す概念に付随する属性やその概念に関連する概念が異なり，コミュニケーションの齟齬が生まれてしまう場合がある．例えば，日本では「鯨」を食べる文化があるが，食用としない文化圏もある．鯨を食用とする文化圏の人と鯨を食用としない文化圏の人が話し合う時に，お互いに食習慣が尊重されていないと感じ，円滑にコミュニケーションが取れない場合がある．このようなコミュニケーションのずれを解消するために，画像検索を用いた対訳ペア間の文化差検出がある． しかしながら，この手法は画像検索の結果に依存するため，主要な画像しか検索されず，文化差のある画像が取得できない場合に正しく機能しない．また，画像に大きな違いのない概念では文化差を検出できない．

そこで，画像には表れにくい文化差を検出するために，概念を単語の分散表現を用いて表し，異言語間で単語の分散表現を対応づけることで文化差を検出する手法を提案する．具体的には，言語ごとのコーパスを用いて各言語の分散表現空間を作成し，敵対的なトレーニングと Procrustes 改良を使用して両空間のアライメントを行う．

本手法の実現にあたり，取り組むべき課題は以下の 2 点である．

異文化間での概念の同一性

日英の単語の分散表現のアライメント後に，対訳関係にある単語が同一の概念かどうか判定するには，類似度の閾値を定める必要がある．基準となる対訳関係の単語ペアを選択し，その全ての単語間の類似度が閾値を超えるように設定する必要がある．

関連概念の分析

対訳関係にある単語ペアの分散表現が類似していたとしても，その概念に関連する概念が異文化間で異なる場合がある．ただし，関連概念とみなす範囲を広げるほど，関連概念間の違いが現れ，誤った文化差を検出する可能性がある．したがって，関連概念を定義するための関連概念の範囲を適切に設定する必要がある．

以下，本研究では 2 章において言語間でのコミュニケーションにおける文化差

を説明し，それに対する現状での文化差検出へのアプローチと問題点を説明する．続いて，3 章において単語分散表現を用い，アライメントを取得する．4 章において，文化差の検出を行う．そして，5 章では 3 章の文化差検出方法と 4 章で導出した最適な閾値を用いて, 文化差を検出できるのか評価を行う．6 章では評価を行い, 7 章では考察を行う．

第2章 異言語間における文化差

本章では, 本研究で取り扱う文化差について, 具体例を用いて説明する. また, 文化差検出の関連研究について記述する.

2.1 異言語間におけるコミュケーションの齟齬

近年, 機械翻訳を用いた多言語でのコミュニケーションが活発化している. しかしながら, 正しく翻訳されているにも関わらず, 文化差によって自分が伝えたい情報をうまく相手に伝えることができないため, コミュニケーションに齟齬が生じる場合がある.

例えば, 日本では「鯨」を食べる文化があるが, 食用としない文化圏もある. 鯨を食用とする文化圏の人と鯨を食用としない文化圏の人が話し合う時に, お互いに食習慣が尊重されていないと感じ, 円滑にコミュニケーションが取れない場合がある.

本研究では, このような異文化コミュニケーションをするときに齟齬を起こすような概念の違いを文化差とし, この文化差を検出することを目的とする.

2.2 関連研究

次に, 多言語コミュニケーションにおける文化差の検出方法に関する既存の研究を示す.

既存の文化差を検出する手法として, 画像特徴量を用いた対訳の文化差検出が存在する[1]. この手法は, 単語で検出される画像の特徴量を用いて文化差の有無を自動判別する. 具体的には, 概念辞書で同一概念に紐付けられている単語を用いる. それらの単語を用いて画像検索を行い, 取得された画像の特徴ベクトルを生成する. 生成されたベクトル間の類似度を計算し, その類似度に基づいた文化差を検出する.

また, Wikipedia を用いた文化差検出手法の提案がある[2]. Wikipedia の記事に含まれる国名・言語名の数を利用し文化差を検出する. 例えば, 異なる 2 つの言語版の Wikipedia において, どちらの記事にも, ある特定の国名・言語名が多い場合には, 各記事は, 同じ内容の説明を行なっている記事であり, 「文化差がない」と判定する. 各言語版の Wikipedia の記事は, 各言語を解する執筆者により書かれる. 逆に, 各言語版の国名・言語名が多い場合は, それぞれの国におけるその言葉

の説明であるため, 各国で違いがある. 各記事において, それぞれの記事の言語の国名や言語名が多い場合には,「文化差」があると判定する.

他にも教師あり学習を利用し, 文化差を検出する手法を提案している[3]. 具体的には言語ごとの Wikipedia コーパスを用いて各言語の分散表現空間を作成し, 教師あり学習によりアライメントをとる. 次に, 概念辞書を用いて, 対象概念に紐付けされている各言語の単語群のベクトルを取得し, 言語ごとの対象概念の統合ベクトルを作成する. その後, 統合ベクトルを用いて文化類似度を算出し文化差の有無を判定する.

第3章 分散表現空間のアライメント

本章では， 文化差を検出するための本研究でのアプローチを説明する． 文化差を検出するために本研究では単語分散表現を用いて文化差の有無を自動判別する手法を提案する．具体的にはまず,Word2Vec で事前学習された日英の単語分散表現空間を取得する． 言語が違うと文法が異なるので， 同じ意味の文であっても対訳関係にある単語の出現する位置が変わってきてしまう． 単語の出現する位置が異なると単語の分散表現も変わってくる． この差異をなくすためにMUSE の教師なし学習を利用し， 異言語間埋め込みを行う．次に， 対訳関係にある単語が同一の概念かどうか判定するために類似度の閾値を定める．基準となる対訳関係の単語ペアを選択し，その全ての単語間の類似度が閾値を超えるように設定する．以下， 図 2 に本研究のアライメントまでのフローチャートを記載する．

3.1 Word2Vec

Word2Vec とは, 文章中の単語を数値ベクトルに変換し, その意味を理解していくという自然言語処理の手法のことである. Word2Vec は, 従来の自然言語処理よりも精度が高く, 単語のベクトル表現をニューラルネットワークの学習を通じて取得することが特徴である.

3.2 分散表現

分散表現（distributed representation）とは, 自然言語処理において, 単語や文章を数値ベクトルで表現する手法である．分散表現として本研究では 300 次元のベクトルを用いる. 分散表現では, 1 つの単語が 1 つのベクトルと対応する. コンピュータは数値の演算しか行えないため, 単語の意味を数値で表現することによって, 意味的な計算を行うことが可能になる. 意味的な計算とは, 主に「近さ」という考え方に基づいている. つまり, 似た意味を持つ単語同士は, ベクトル同士の cos 類似度が大きくなるように配置される.

例えば,「King」と「man」の間のベクトルは.「Queen」と「woman」の間のベクトルと近似できる. この性質により,

「King」－「man」＋「Queen」＝「woman」となるような意味の差や和などの等式を近似できるようになる.

3.3 機械学習

機械学習（Machine Learning）とは, コンピューターにデータを学習させて, アルゴリズムに基づいて分析させる手法である. 機械学習では, アルゴリズムやモデルを使用して, データからパターンや関係性を学習する.

機械学習には「教師あり学習(Supervised Learning)」,「教師なし学習(Unsupervised Learning)」,「強化学習(Reinforcement Learning)」の 3 種類が存在する.

3.3.1 教師あり学習

教師あり学習は, コンピューターにラベル付きのデータを与え, 正しい出力を教えることで学習させる方法である. 訓練データと呼ばれるデータを使用し, 入力と出力の関係を学習し, 新データに対して正確な出力を生成できるようになる.

例えば, コンピューターに, 画像が「犬」なのか「猫」なのかを判別させたい場合,「犬」や「猫」とすでにラベルが付いた訓練データを与えることによってコンピューターに「犬」と「猫」の特徴を学習させる. その結果, ラベルのない画像を入力しても正しくコンピューターは正しく「犬」や「猫」を判別できるようになる.

3.3.2 教師なし学習

教師なし学習では, コンピューターにラベルの付いていないデータを与えることで, そのデータの特徴やパターンを学習させる. 教師なし学習の場合, 正しい出力が定義されていないため, 入力データがどのポイントやグループに近いかを判別するように訓練する. 本研究で教師なし学習を使用した目的は, 日本語と英語のすべての可能な対応を考慮するためである. 例えば, 日本語の「赤」が英語の「red」に対応するかもしれないが, 「薄赤」が「red」に対応するかもしれない. もし, 「薄赤」が「red」に対応するなら「薄赤」と「red」に付随する属性やその概念に関連する概念にずれが生じ, その影響によりコミュニケーションの齟齬が生まれてしまう場合がある.

3.3.3 強化学習

強化学習は, コンピューターの出力に対して点数を与え, 点数の高い出力を学習させる方法である. 良い結果には報酬を与えることで, 機械により良い出力を学習させることができる.

3.4 異言語間単語埋め込み(MUSE)

MUSE (Multilingual Unsupervised and Supervised Embeddings)は, 異言語間での単語や埋め込みを学習するための手法である. これは, 機械翻訳やクロス言語情報検索などの自然言語処理タスクにおいて, 異なる言語間での単語や文の翻訳の関連付けを行うことができる. つまり, 異なる言語の分散表現空間が同じ空間に埋め込まれるように学習することである.

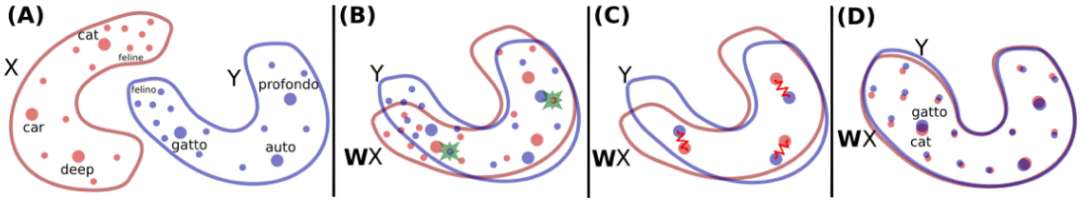

図 1 : MUSE [5]

図 1 の MUSE の説明をする. (A) 赤で示される英語の単語 (X) と青で示される日本語の単語(画像上ではイタリア語である) (Y) の 2 つの埋め込み分布があり, これらのアライメントをとる. 各ドットはその空間の単語を表している. ドットのサイズは言語のトレーニングコーパス内での単語の出現頻度に比例している. (B) 敵対的トレーニングを使用して, 2 つの空間のアライメントをとる回転行列Wを学習する. 緑の星は, 2 つの単語埋め込みが同じ空間から来たかどうかを判別するために判別器に供給されるランダムに選択された単語である. (C) マッピングWは Procrustes によりさらに精度が上がる. (D) 最後に, マッピングWと距離メトリクス (CSLS と呼ばれる) を使用して翻訳する.

言語が異なると, 同じ意味で対訳となっている文章であっても, 対訳となる単語の出現位置が変わる. 単語の出現位置が変わると同じ意味の単語であっても分散表現に出現する位置も変わってしまう. よって, 言語の違いによる分散表現のずれを補正する必要がある. このタスクを異言語間単語埋め込みという. 本研究では Facebook が公開しているライブラリである MUSE の教師なし学習を使用し, 英語の分散表現空間を日本語の分散表現空間に近づける形で異言語間単語埋め込みを行った. つまり, 英語がソース(source), 日本語がターゲット(target)として, ソース言語がターゲット言語に合わせる形で研究を行なった.

3.4.1 教師なし学習の概要

本研究では,Word2Vec によりトレーニングされたベクトル空間が 2 つある．日英間でアライメントをとる時の回転行列，つまり，ソースとターゲットの空間の間に線形マッピングWを学習する．

$$W *= \ argmin\|WX - Y\|_{\mathrm{F}} \tag{1}$$

ここでdは埋め込みの次元，$M_d(\mathbb{R})$は実数の$d \times d$行列の空間，Xと Yは$d \times n$のサイズで，平行語彙内の単語の埋め込みを含む 2 つの整列をした行列である．ソース言語 s の翻訳 t は，$t = arg\ max_t\ cos(Wx_s, y_t)$と定義される．

さらに精度を上げるため，行列 W に直行性制約を課せる．つまり，Wの行ベクトル同士の内積が 0 になるようにする．方程式（1）は Procrustes 問題という 2 つの行列間で最も適切な対応を見つける問題に帰着する．また，行列を効果的に分解する手法の一つである特異値分解（SVD）を利用し，任意の行列を特定の構造に分解することができる．最後に，特異値分解から得られる式の解を使う．

$$W *= \ argmin\|WX - Y\|_{\mathrm{F}} = \ UV^T\ , with\ U\Sigma V^T\ = \ SVD(YX^T) \tag{2}$$

異言語間埋め込みをするために，このマッピング Wを学習する．まず，敵対的な基準を使用して Wの初期の近似を学習する．つまり，異言語間での類似性を考慮して，初期の Wを学習する.次に，最も意味の近い単語ペアをアンカーポイントとして選択する．選択されたアンカーポイントを用いて，Procrustes 解法を適用し，マッピング行列 Wを調整する．最後に，空間の距離を変更して，これにより単語の表現がより適切になり，特に出現頻度の低い単語に対する精度の向上が行われる．次に，これらの手順の詳細を説明する．

3.4.2 教師なし学習のドメイン敵対的設定

この章では，教師なし学習で Wを学習するためのドメイン敵対的設定を説明する．$\mathcal{X} = \{x_1, \ldots, x_n\}$及び$\mathcal{Y} = \{y_1, \ldots, y_n\}$を，それぞれソース言語とターゲット言語からくる$n$個と$m$個の単語埋め込みを行う．この方法では，まずモデルを訓練する．$W\mathcal{X} = \{W_{x_1}, \ldots, W_{x_n}\}$及び$\mathcal{Y}$からランダムにサンプリングされた要素を区別できるように訓練する．ソース言語の単語埋め込み集合と，ターゲット言

語の単語埋め込み集合からランダムにサンプリングされた要素を取り出す． 取り出した要素がどちらの言語からきたものかを区別できるように， マッピングモデルを訓練する． マッピング行列Wを学習するために， 敵対的な学習設定を行う． これは， ソース言語からきた埋め込みとターゲット言語からきた埋め込みを分離するようにモデルを訓練する．

Discriminator objective

判別器のパラメータをθ_Dとする．ベクトルzが判別器によってソース埋め込みのマッピングである確率$P_{\theta_D}(source = 1|z)$とする． 判別器の損失は以下のように書くことができる．

$$\mathcal{L}_D(\theta_D|W) = -\frac{1}{n}\sum_{i=1}^{n}\log P_{\theta_D}(source = 1|Wx_i) - \frac{1}{m}\sum_{i=1}^{m}\log P_{\theta_D}(source = 0|y_i) \quad (3)$$

Mapping objective

教師なし学習では，Wは判別器が埋め込みを正確に予測できないように訓練される．

$$\mathcal{L}_W(W|\theta_D) = -\frac{1}{n}\sum_{i=1}^{n}\log P_{\theta_D}(source = 0|Wx_i) - \frac{1}{m}\sum_{i=1}^{m}\log P_{\theta_D}(source = 1|y_i) \quad (4)$$

Learning algorithm

モデルを訓練するために, 敵対的ネットワークの訓練を行う. 各入力サンプルに対して, 判別器とマッピング行列Wは, それぞれ$\mathcal{L}_D$と$\mathcal{L}_W$を最小化することでマッピング行列Wが正しいマッピングを学習できるようになる．

3.4.3 正確性の向上手順

次に， 敵対的訓練で学習した「W」を用いて合成整列語彙を構築する． 具体的には， 最も頻度の高い単語を考慮し， 相互最近傍のみを保持する． これにより，高品質の辞書が得られる．その後，生成した辞書に対して(2)を適用する．

3.5 アライメント手順

この節では, Wikipedia の文章から生成された分散表現空間を MUSE の教師なし学習を用い, 日英間の単語ペアのアライメントを行う. フローを簡単に理解するために図 2 のフローチャートとともに, 具体的な説明を行う.

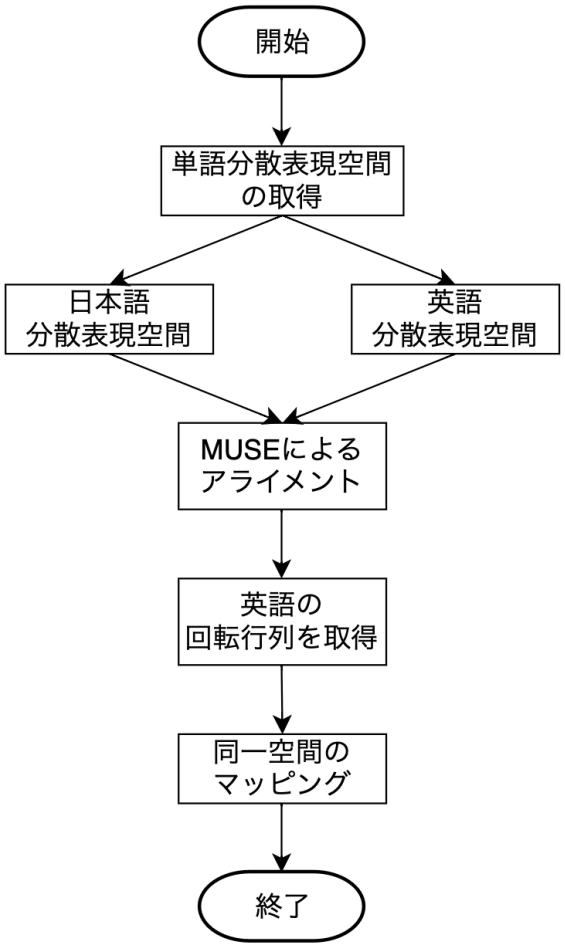

図 2 ： 異言語分散表現を用いたアライメント

まず, Wikipedia の文章を Word2vec で学習し, 表現力の高い 300 次元の単語分散表現空間の取得を行う. 次に, 日本語分散表現空間と英語分散表現空間を MUSE の教師なし学習を用いてアライメントをとる. この時に, 日本語分散表現空間をターゲットとしてソースである英語分散表現空間を回転させる. この時に英語の回転行列W(2)が生成されるので, 英語分散表現空間を日本語分散表現

空間に回転させ，同一空間へのマッピングを行い，終了する．

第4章 文化差抽出

この章では文化差を抽出するための全体の流れを図 3 のフローチャートを使用し, 具体的に説明する.

前章では, 日本語分散表現空間を中心として, 英語分散表現空間を回転させ, 同一空間へのマッピングを行った. 英語の回転させた後のベクトルと日本語のベクトルを用い, 類似度行列の作成を行う. 次に, 類似度行列の対角成分が高くなっていることを確認し, k-近傍法により $k \leq 5$ で対訳ペアが発見できれば「文化差なし」, 発見できなければ「文化差あり候補」の抽出を行う. 最後に「文化差あり候補」がスワデシュリストの閾値よりも高ければ,「 文化差なし」, 閾値よりも低ければ「文化差あり」と判定し, 終了する.

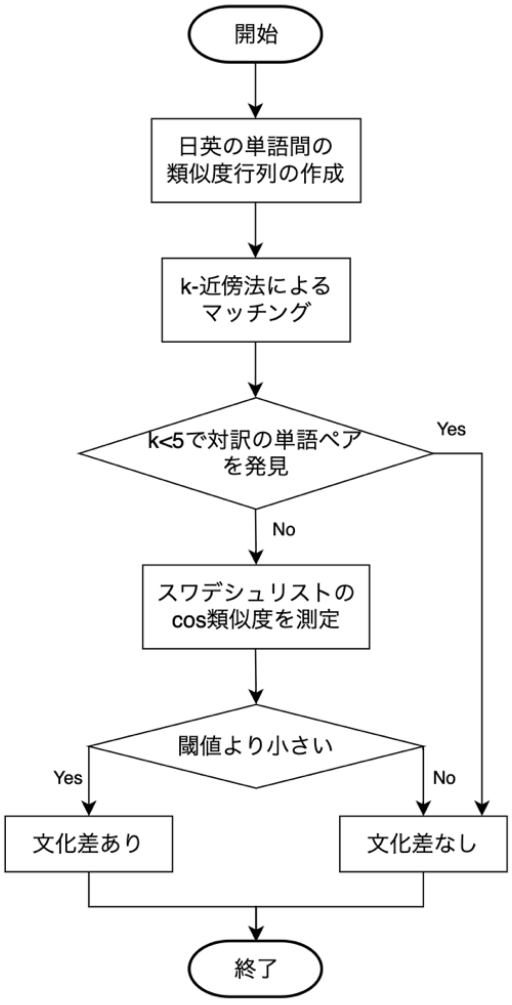

図 3 : 文化差抽出

4.1 同一概念の同定

この節では k-近傍法による対訳ペアの同一性チェックを説明する．まず, MUSE の教師なし学習でアライメントを行なった. 次に, 正しくアライメントがとれていることを確認するために, 類似度行列の対角成分の cos 類似度が高くなっているかを可視化し, 判断する. ここで対角成分の類似度が高くなっていなければ, アライメントが正しくとれていないことになるので, k-近傍マッチングで誤った絞り込みを行ってしまう．誤った絞り込みを行うと, 正しく文化差を検出できなくなる. 正確な類似度行列を用いて, k-近傍法によるマッチングで絞り込みを行うことにより, アライメントが取れていない単語を抽出することができる. そのアライメントが取れていない指標を, $k \geq 5$ とする. つまり, $k \geq 5$ で発見された単語のペアを「文化差がある候補」としてピックアップする.

4.2 文化差候補の絞り込み

スワデシュリストとは, 言語間の類似度を定量的に評価するために使用される, 基本的な語彙を収集したリストのことである. このスワデシュリストを用いて, 基準とした閾値による絞り込みの説明を行う. スワデシュリストの cos 類似度を全て測り, 最も低かった数値を閾値とする. 最後に, 同一単語で $k \geq 5$ かつ, 閾値を超えていないペアを抽出する. さらに, 抽出された単語のペアの cos 類似度上位 10 件を比較し, それらの単語にどのような属性の違いがあるのかを分析する.

第5章 文化差抽出実験

5.1 学習データ

本手法では Word2Vec で学習された単語分散表現空間を用いる. Wikipedia から学習された大規模な 1,854 個のオブジェクト概念（THINGS）とシュワデスリストをアライメントの対象とする. 1,854 個のオブジェクトとは, まず, 既存の単語データベースから, 具体的でイメージ可能なオブジェクトの概念を表す名詞のリストを収集する. 次に, 各名詞に, この名詞の意味を表す 1 つまたは複数の一意の WordNet 識別子（「Synset」）を割り当てることで語義の曖昧さを解消した. これにより, 同義語を削除したり, 複数の意味を持つ名詞を識別したりできるようになる. 最後に, 全ての新セットの代表的な画像を選択し, それらが人間の被験者によってどれだけ一貫して命名されているかをテストすることにより, 日常言語での使用に一致するシンセットのサブセットを特定した. スワデシュリストとは,「基礎語彙」を集めたさまざまなリストのうちの一種である. これは, 言語統計学, 言語間の近縁度の量的な見積もりに用いられる.

5.2 教師なし学習のアライメント

この節では 3.8.2 節から 3.8.4 節のアルゴリズムに従い, THINGS とスワデシュリストのアライメントを取る. その際に, ターゲットである日本語のベクトルを固定し, ソースの英語のベクトルを（2）の式を利用し, 回転させる. 以下の図 4 が回転させた後のベクトルである. 赤色のポイントが英語の単語で青色のポイントが日本語の単語になっている. 図 4 だけではアライメントを正しくとれているかを確認できないため, 類似度行列（図 5）を作成することにした.

5.3 類似度行列

類似度行列(similarity matrix)とは, データセット内の各要素間の類似性を表す行列である. データセット内の各要素は 300 次元のベクトルで表されており, それらの cos 類似度を計算した. 横軸は日本語で縦軸は英語になっており, アルファベット順になっている. 最後の要素にスワデシュリストを配置している. 類似度行列の対角成分の類似度が高くなっており, 正確にアライメントが取れていることが確認できる. 次に, MUSE での教師なし学習のアライメントの

性能を評価することと，アライメントが取れていない閾値としてk-近傍法を計算することにした.

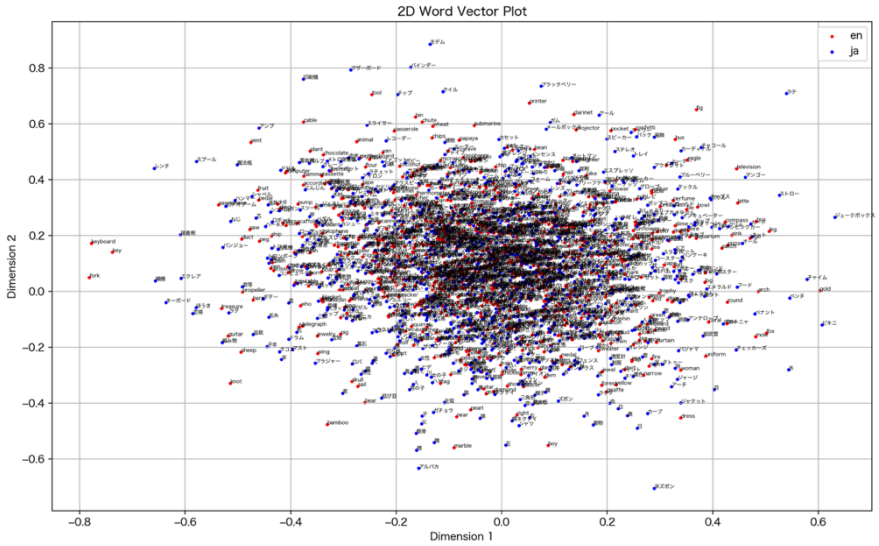

図４： ベクトルのプロット

図５： 類似度行列

5.4 k-近傍マッチング

k-近傍法(k-Nearest Neighbors)とは，機械学習の分類及び開基の手法の一つである．K-近傍法は，与えられたデータポイントの近くにあるトレーニングデータポイントのうち，最も近い k 個の点を見つけ，その点を取り，あるデータにマッチしている最も近い k 個の点を見つけ，その点を取り，あるデータにマッチしているかを分類する．今の場合，日本語と英語の分散表現空間からアライメントされた空間で対訳が k-近傍にある場合正しく一致しているとみなす．例えば，「apple」という英単語が「りんご」を見つけるのに apple から最も近い単語を順番に探していく．k=1 ならば apple に一番近い単語を 1 つ探索する．k=2 ならば apple に 2 番目に近い単語を探索するということである．図 6 は k=1 から k=9 までの単語を探して，見つかった割合を縦軸に示した．結果が，k=1：0.00%，k=2：0.00%，k=3：7.89%，k=4：69.30%，k=5：95.59%，k=6：99.44%，k=7:99.91%，k=8：100%である．k=5 までに，ほとんどの単語が発見されているため，MUSE のアライメントは正確に行われたと言える．

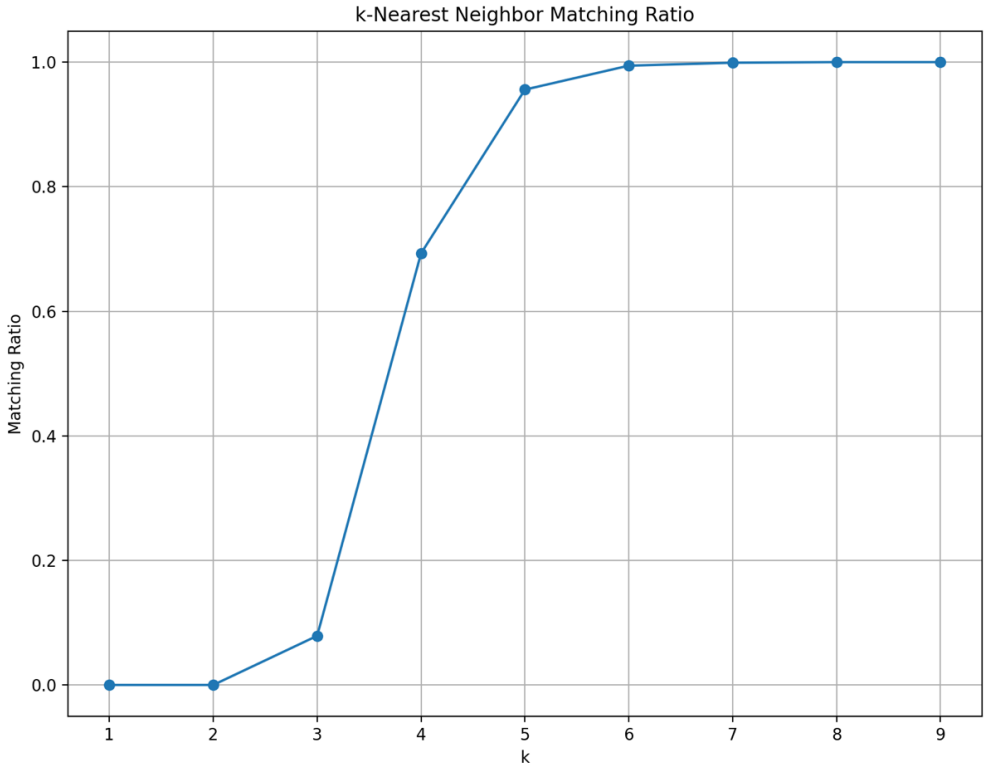

図 6 ： k-近傍法

表 1：k=5, 6, 7 で抽出された文化差候補リスト

日本語	英語	日本語	英語
たる	barrel	米	rice
豆	bean	ルート	root
ブーツ	boot	スコップ	scoop
ブレース	brace	得点板	scoreboard
充電器	charger	アザラシ	seal
顎	chin	船	ship
柱	column	ナメクジ	slug
紙吹雪	confetti	拍車	spur
コード	cord	杭	stake
ベビーベッド	crib	魔法瓶	thermos
うちわ	fan	ダニ	tick
イチジク	fig	亀	turtle
フォーク	fork	骨壺	urn
菌類	fungus	通気孔	vent
帽子	hat	歩行者	walker
蛍光ペン	highlighter	ヤク	yak
アイロン	iron	クランク	crank
ジャー	jar	取っ手	handle
鍵	key	生垣	hedge
キウイ	kiwi	丸太	log
大理石	marble	クワッド	quad
モグラ	mole	受け皿	saucer
玉ねぎ	onion		

また，k=5 から k=7 までにいくつかの単語が見つかっていないため，どういう単語が見つからなかったのかを確認し分析する． 上記の表 1 は k-近傍法を計算したときに k=5, k=6, k=7 でも見つけられなかった単語のリストである．k=6 で見つからなかった単語は(クランク crank)，(取っ手 handle)，(生垣 hedge)，(丸太 log)，(受け皿 saucer)． k=7 で見つからなかった単語は(クワッド quad)．k=5 で見つからなかった単語はそれ以外の単語である．

表 1 の中には多義語の可能性を考慮し，それを除外する必要がある．例えば，

表 2 ： 同義語

日本語	英語	cos類似度
豆	bean	0.4777
ベビーベッド	crib	0.3749
菌類	fungus	0.5368
亀	turtle	0.6721
通気孔	vent	0.3859

（生垣 hedge）のペア，hedge は生垣の意味があるが，リスクヘッジなどに使われる「防止策」という意味もある． このように， アライメントを取った際に，文化差ではなく，単語ペアの上位 10 件を調べ，単なる多義語が影響し，k-近傍でうまく取れなかったペアは除去する．上記のリストで（たる barrel）,（ブーツ boot）・（ブレース brace）・（充電器 charger）・（顎 chin）・（柱 column）・（紙吹雪 confetti）・（コード cord）・（うちわ fan）・（フォーク fork）・（帽子 hat）・（蛍光ペン highlighter）・（アイロン iron）・（ジャー jar）・（鍵 key）・（キウイ kiwi）・（大理石 marble）・（モグラ mole）・（玉ねぎ onion）・（米 rice）・（ルート root）・（スコップ scoop）・（得点板 scoreboard）・（アザラシ seal）・（船 ship）・（ナメクジ slug）・（拍車 spur）・（杭 stake）・（魔法瓶 thermos）・（ダニ tick）・（骨壷 urn）・（歩行者 walker）・（ヤク yak）・（クランク crank）・（取っ手 handle）・（生垣 hedge）・（丸太 log）・（クワッド quad）・（受け皿 saucer）の 41 ペアを省く．多義語の影響により 41 ペアをリストから削除した．そして，リストに残った 6 ペアは同義語なので，cos 類似度と共に表 2 に載せる．

5.5 閾値の設定

閾値の設定のために, スワデシュリストを基準とした絞り込みを行う．スワデシュリストの cos 類似度を全て測り, 最も低かった数値を閾値とする. その際に, スワデシュリストでも多義語の影響があるため, 多義語を考慮し, 同義語のペア

を閾値として使用する. 同義語は前節と同様に cos 類似度上位 10 件を取得し, 判断を行う.

表 3 にスワデシュリストの中で cos 類似度が低い 6 件を示す. この中で閾値の候補として最も低い値（日 sun）を採用したいが多義語の影響により, アライメントが取れていないだけだった. また, （月 moon）・（丸い round）・（私 I）・（灰 ash）が多義語の影響を受けていたので, （人 human）を閾値として採用することにする.

表 3 : スワデシュリストの閾値

日本語	英語	cos類似度
日	sun	0.3208
月	moon	0.3708
丸い	round	0.3929
私	I	0.4245
灰	ash	0.4307
人	human	0.4444

5.6 文化差抽出

この節では, 文化差の有無を判断する. 5.4 節の k-近傍法で文化差がある候補を抽出し, 5.5 節で閾値の設定を行なった. ここからは, 文化差がある候補に閾値を設け, 足切りを行い, 文化差の有無を出力する.

5.5 節で閾値が cos 類似度 0.444 と設定したので, 表 2 から cos 類似度 0.444 より小さい単語のペアを探す. その結果, （ベビーベッド crib）, （イチジク fig）, （通気孔 vent）を「文化差がある」という対象として, 抽出することにする.

第6章 結果

この章では文化差抽出実験をした結果を示す．特に, 前章で「文化差がある」という対象として抽出した（ベビーベッド crib），（イチジク fig），（通気孔 vent）について cos 類似度上位 10 件を取得した. 表 4, 表 5, 表 6 は上から順に cos 類似度が高く日本語と英語それぞれ 10 件並んでいる．

まず,（ベビーベッド crib）のペア, 表 4 の評価を行う．日本語分散表現空間には「##ベビーベッド##」,「おむつ交換台」,「 トイレ」,「##休憩所##」,「##便所##」,「##多目的トイレ##」,「##ベビーチェア##」など「赤ちゃん」に関する単語が並んでいる．一方で, 英語分散表現空間では,「crib」に直接関する単語は見つけられない．

2 つ目に,（イチジク fig）のペア, 表 5 の評価を行う．日本語分散表現空間では「果実」,「果物」,「##イチジク##」など「イチジク」に直接関係する単語が

表 4 : ベビーベッドの関連概念(10 件)

日本語	英語
ベビーベッド	crib
##ベビーベッド##	drawer
おむつ交換台	sowing
オストメイト	slicing
##公衆電話##	shovel
トイレ	stagger
##休憩所##	hearth
##便所##	pronoun
##郵便ポスト##	mutilation
##多目的トイレ##	fry
##ベビーチェア##	barge

表 5 : イチジクの関連概念(10 件)

日本語	英語
イチジク	fig
果実	zool
##ポーポー##	pls
ブドウ	naturkundemuseum
果物	biodiversitylibrary
##イチジク##	guggenheim
##オプンティア##	supp.
##ロサ・ギガンティア##	biol
タコノキ	figs
ドライフルーツ	brill
##ネクタリン##	Lundae

表 6 : 通気孔の関連概念(10 件)

日本語	英語
通気孔	vent
空気弁	vents
煙出し	chute
エジョクションポート	pieds
通気口	mains
##エアシャワー##	Vent
##原子炉建屋##	flute
##給水管##	terre
アプリケータ	dans
##開口部##	orgue
焚口	mixte

その他にも,「##ポーポー##」,「ブドウ」,「##オプンティア##」,「##ロサ・ギガンティア##」,「タコノキ」,「ドライフルーツ」,「##ネクタリン##」など, 果物に関する単語が並んでいる. 一方で, 英語分散表現空間では,「figs」という「fig」に直接関係する単語が並んでいる. その他には,「zool」,「naturkundemuseum」,「biodiversitylibrary」, など「fig」が置かれている施設が上位にきていた.

3 つ目に,（通気孔 vent）のペア, 表 6 の評価を行う. 日本語分散表現空間では「空気弁」,「煙出し」,「エジョクションポート」,「通気口」,「###原子炉建屋#」,「開口部」,「焚口」など「通気孔」に直接関する単語が並んでいる. 一方で, 英語分散表現空間では,「vents」,「chute」,「Vent」など「vent」に直接関する単語が並んでいる.

第7章 考察

7.1 文化差ありと判定した単語の考察

まず,（ベビーベッドcrib）のペア,表 4 の考察を行う. 日本語分散表現空間では最も cos 類似度が高いところに「##ベビーベッド##」がきている. その他にも「おむつ交換台」,「トイレ」,「##休憩所##」,「##便所##」,「##多目的トイレ##」,「##ベビーチェア##」など「赤ちゃん」に関する単語が並んでいる. 一方で, 英語分散表現空間では, ベビーベッドに直接関連する単語が並んでいない. しかし,「stagger」や「hearth」などベビーベッドを間接的に連想させる単語を発見することができた. 表 4 から考察するに, 日本では「ベビーベッド」そのものをイメージするが, 英語圏では, ベビーベッドが「揺れている」ことをイメージすると考えられる.

2 つ目に,（イチジク fig）のペア, 表 5 の考察を行う. 日本語分散表現空間では「果実」,「果物」,「##イチジク##」など「イチジク」に直接関係する単語が並んでいる. その他にも,「##ポーポー##」,「ブドウ」,「##オプンティア##」,「##ロサ・ギガンティア##」,「タコノキ」,「ドライフルーツ」,「##ネクタリン##」など, 果物に関する単語が並んでおり,「イチジク」に間接的に連想できる単語が並んでいる. 一方で, 英語分散表現空間では,「figs」という「fig」に直接関係する単語が並んでいる. その他には,「zool」,「naturkundemuseum」,「biodiversitylibrary」, など「fig」が置かれている施設を発見することができた. これらから, 日本では「イチジク」そのものをイメージするが, 英語圏では,「イチジク」そのものもイメージするが, イチジクが置かれている「施設」をイメージすると考えられる.

3 つ目に,（通気孔 vent）のペア, 表 6 の考察を行う. 日本語分散表現空間では「空気弁」,「煙出し」,「エジョクションポート」,「通気口」,「###原子炉建屋#」,「開口部」,「焚口」など「通気孔」に直接関する単語が並んでいる. 一方で, 英語分散表現空間では,「vents」,「chute」,「Vent」など「vent」に直接関する単語が並んでいる. このことから日本では「煙出し」や「エジョクションポート」が上位にきている点,「煙を出す」イメージが強く, 英語圏では「射水路」が上位にきている点,「水を出す」イメージが強いと考えられる.

4 つ目に,（クジラ whale)の文化差が存在すると予想していたため, 調査を行

なったが， cos 類似度は 0.709 と閾値の 0.444 よりも高くなり，「文化差なし」と判断した．また，（クジラ whale)の cos 類似度上位 10 件を探した時に，日本語分散表現空間と英語分散表現空間では取得してくる単語が大きく異なるのではないかとも考えたが，両方ともにイルカや別のクジラなどが発見され，文化差はないと考えられる．

7.2 問題点

この節では前節で考察したことを踏まえながら，問題点を考える．表 4，表 5，表 6 において，日本語の近縁単語は直接的で正しい場合が多いが， 英語の近縁単語は明らかに関係のない単語も含まれていることが多く，正しく文化差を検出できているとは言い難い．

英語の近縁単語が明らかに関係のない単語を含んでいることが多かった理由として，今回用いたデータでは，日本語の単語数と比べて英語の単語数が少なかったことが考えられる． よって，上位 10 件を探しても，全く関係のない単語を取得してくる可能性が高い．このことから，英語の単語数を増やすことで，より高い精度で，文化差を検出できると考える．

第8章 おわりに

多言語コミュニケーションにおける文化差を解消するために,本研究では単語分散表現から得られるベクトルを使用し,日英間でアライメントをとることによって比較するアプローチを提案してきた.本研究の貢献は以下のとおりである.

異文化間での概念の同一性

k-近傍法により，k=5 から k=7 の間で未発見の単語かつ同義語の単語ペアは（豆 bean)・(ベビーベッド crib)・(菌類 fungus)・(亀 turtle)・(通気孔 vent）だった．次に，文化差がない基準としてスワデシュリストの cos 類似度を測った．（人 human）の cos 類似度 0.444 が最も低かったので，この数値を閾値として扱う．この閾値より低い単語のペアは（イチジク fig），（ベビーベッド crib），（通気孔 vent）だった．また，概要で述べた（クジラ whale）の cos 類似度は 0.709 と閾値を超えた．上位 10 件を探すと，日本語と英語ともにイルカや別のクジラが発見され，「文化差あり」とは判断し難い結果となった.

関連概念の分析

関連概念の範囲として単語の cos 類似度上位 10 個に設定した．k-近傍法で上位 10 件を取得した結果，（イチジク fig），（ベビーベッド crib），（通気孔 vent）は直接的にあるいは間接的に該当の単語に一致していた．しかし，英語分散表現では明らかに関係のない単語も散見された.

文化差を検出するために単語分散表現空間を使用し，日英間でアライメントを行った．次に，k-近傍法と閾値の設定により文化差がある単語ペアの候補を抽出し，cos 類似度上位 10 件を比較し，文化差を検出することができた．しかし，英語分散表現空間の上位 10 件の中には明らかに関係の無い単語も取得してしまっているため，文化差を検出するにはさらに精度を高める必要がある.

謝辞

本研究を行うにあたり，ご指導していただいた村上陽平教授に深く感謝を申し上げます.

参考文献

[1] 西村一球: 画像特徴量を用いた対訳の文化差検出,ヒューマンインタフェース学会論文誌 23巻, 2号, p.145-152.(2021)

[2] 宮部真衣, 吉野孝:Wikipedia を用いた文化差検出手法の提案,情報処理学会論文誌, マルチメディア, 分散, 協調とモバイル（DICOMO2011）, シンポジウム, pp30-36(2011-07).

[3] 大井也史: 異言語間の分散表現を用いた文化差検出, pp1-2(2020)

[4] Genji Kawakita, Ariel Zeleznikow-Johnston, Ken Takeda, Naotsugu Tsuchiya and Masafumi Oizumi1: Is my “red” your “red”? :Unsupervised alignment of qualia structures via optimal transport, pp1-5(2023)

[5] Conneau, Alexis and Lample, Guillaume and Ranzato, Marc'Aurelio and Denoyer, Ludovic and Jégou, Hervé: Word Translation Without Parallel Data, arXiv preprint arXiv:1710.04087,(2017)