卒業論文

異言語間の分散表現を用いた
文化固有単語の対訳生成

指導教官　　村上 陽平 准教授

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
2600180416-1

ZHONG Ganghui

2021年度（秋学期）卒業研究3（CH）
令和4年1月31日

異言語間の分散表現を用いた文化固有単語の対訳生成

ZHONG Ganghui

内容梗概

近年，ニューラルネットワークの発展に伴い，機械翻訳の翻訳精度が向上し，多くの形式的な文書（論文，法律文書等)は正しく翻訳することが可能になっている．しかしながら，生活や宗教など，文化的に独自の色合いの強い文章は，文化，思考と習慣に基づいて形成されているため，その言語にしか存在しない単語が含まれる場合がある．例えば，中国語で「汤圆」とは，中国伝統的なお祭りである「元宵節」に食べるもので，もち米の粉などで作られたゴマの入った球状で，主に似て食べるものという意味である．日本語には形で似たような食べ物はあるが，「元宵節」というお祭りで食するという文脈を表す単語が存在しない．そのため，「正月十五，家家户户吃汤圆，挂灯笼.」を機械翻訳で翻訳すると「朔望月の最初の 15 日目に，すべての家族が餃子を食べ，提灯を吊るします.」という翻訳結果が生成され，「汤圆」は「餃子」と誤訳されている．このような文章を正確に翻訳するには，文化固有単語の対訳の生成が必要不可欠である．

そこで，本研究では，異言語間の単語分散表現を用いて，原言語の文化固有の単語に対応する，目的言語の類似語を対訳として生成する．具体的には，日中それぞれの Wikipedia コーパスから各言語の分散表現空間を生成し，対訳辞書データに基づいて，両空間のアライメントを行う．次に，中国語の固有単語を日本語のベクトル空間に写像し，日本語の類義語を探すことで，固有単語の対訳を生成する．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．

類義語の組み合わせ

日中の分散表現空間の対応付けを行ったとしても，中国語の単語ベクトルと極めて近似する単語ベクトルを日本語の分散表現空間で得ることは難しい．対応する複数の日本語の類義語の中から中国語の単語を説明するのに有用な類義語の全組み合わせで新しいペアを生成して説明する必要がある．

類義語と差分の組み合わせ

中国語の単語で説明するのに類義語の全組み合わせで取得したペアでは，あまり意味を表現できない可能性がある．そのため，類義語ベクトルと中国語の単語ベクトルを別々に引き算で取得した最も近い日本語の単語と，それに対応する日本語の類義語を組み合わせ，新たに得られ

たペアを用いて中国語の単語を再度解釈する.

前者の課題に対しては，分散表現空間のアライメント後に得られた日本語の 20 件の類似概念から，中国語の単語を最も良く表す組み合わせを発見する．具体的には，単語の組み合わせごとに単語ベクトルの平均を計算し，その平均ベクトルと中国語の単語ベクトルのコサイン類似度を計算し，最も類似する単語の組み合わせを選択する.

後者の課題に対しては，20 個の日本語類義語ベクトルと中国語の単語ベクトルをそれぞれ減算し，新たな差分ベクトルにより最も近い日本語単語を求め，その単語と対応する日本語の類義語との間で新たなペアが形成される．取得したペアの和ベクトルと中国語の単語ベクトルのコサイン類似度を計算し，最も類似する単語の組み合わせを選択する.

Wikipedia から選択した中国語固有の単語を 10 件に対して，本提案手法を適用し，選択された日本語の類似概念の組み合わせのうち，人手で最適と判断した組み合わせの順位を用いて評価を行い，提案手法の有効性を検証した．本研究の貢献は以下の通りである.

類義語の組み合わせ

アライメントを行った日中の分散表現空間を用いて，中国語の文化固有単語に類似する日本語の単語 20 件を取得し，全単語ペアの平均ベクトルの類似度よって，順序付けを行った．提案手法により，単純に日本語の類似単語を対訳とするよりも，平均逆ランク(MRR)が 12.78%，平均適合率の平均（MAP）が約 97.15%向上した.

類義語と差分の組み合わせ

類義語の組み合わせだけでなく，類義語と差分の組み合わせも用いることで，中国語の文化固有単語の説明を生成した．提案手法により，単純に日本語の類似単語を対訳とする場合と比べて，平均逆ランク(MRR)が 72.43%，平均適合率の平均（MAP）が 18.53%向上した.

Translation of culture-specific words using cross-lingual distributed representation

ZHONG Ganghui

Abstract

Recently, with the development of neural networks, the translation accuracy of machine translation has improved and many formal documents (theses, legal documents, etc.) can be translated correctly. However, texts with strong cultural characteristics, such as life and religion, are formed based on culture, thoughts, customs, and may contain words that exist only in that language. For example, the Chinese word "汤圆" refers to a type of food eaten during the Lantern Festival, a traditional Chinese festival, which is a sesame-filled ball made of glutinous rice flour, etc. In Japanese, there are foods that are similar in shape, but there is no word that indicates the context in which they are eaten during the "Lantern Festival". Therefore, "正月十五，家家户吃汤圆，挂灯笼. would be translated as "On the first 15 days of the first month, all families eat dumplings and hang lanterns. " and "汤圆" is mistranslated as "dumplings". In order to translate such sentences accurately, it is essential to generate a bilingual translation of culture-specific words.

In this paper, we used cross-linguistic word variance expressions to generate bilingual similar words in the target language that correspond to culture-specific words in the source language. Specifically, we generated a distributed representation space for each language from the Wikipedia corpus of Japanese and Chinese, and aligned the two spaces based on bilingual dictionary data. Then, we mapped the Chinese Eigenwords, a real-valued vector embedding associated with a word, to the Japanese vector space and search for Japanese synonyms to generate the bilingual dictionary of the Eigenwords. In the realization of this technique, the following two points should be tackled.

Synonym combination

Even if we map the Japanese and Chinese distributed representation spaces, it is difficult to obtain a word vector in the Japanese distributed representation space that is very close to the Chinese word vector. It is necessary to generate a new pair from all the combinations of synonyms that are useful for explaining the Chinese word among several corresponding Japanese synonyms.

Synonyms and differential combination

The pairs obtained from all the combinations of synonyms may not be able to express the meaning of the Chinese words. Therefore, we combine the closest Japanese word obtained by subtracting the synonym vector and the Chinese word vector separately with its corresponding Japanese synonym, and reinterpret the Chinese word using the newly obtained pair.

For the former task, we find the combination that best represents the Chinese word from the 20 similar concepts in Japanese obtained after the alignment of the distributed representation space. Specifically, we calculate the average of the word vectors for each word combination, calculate the cosine similarity between the average vector and the Chinese word vector, and select the most similar word combination.

For the latter task, the 20 Japanese synonym vectors are subtracted from each of the Chinese word vectors and the new difference vector is used to find the closest Japanese word, and a new pair is formed between that word and the corresponding Japanese synonym. The cosine similarity between the sum vector of the obtained pairs and the Chinese word vector is calculated, and the most similar word combination is selected.

We applied the proposed method to 10 Chinese-specific words selected from Wikipedia, and evaluated the effectiveness of the proposed method by using the rankings of the combinations of the selected Japanese similar concepts that we judged to be the best manually. The contributions of this study are as follows.

Synonym combination

Using the aligned Japanese-Chinese distributed representation space, we obtained 20 Japanese words that are similar to Chinese culture-specific words, and ordered them according to the similarity of the mean vector of all word pairs. The proposed method improves the Mean Reverse Rank (MRR) by about 12.78% and the Mean Average Percentage of Fitting (MAP) by about 97.15% compared with the simple bilingualization of similar Japanese words.

Synonyms and differential combination

By using not only the combination of synonyms but also the combination of synonyms and differences, we generated descriptions of culture-specific words in Chinese. The proposed method improves the mean reverse rank (MRR) by 72.43% and the mean average percentage of fit (MAP) by 18.53% compared to the simple translation of similar words in Japanese.

異言語間の分散表現を用いた文化固有単語の対訳生成

目次

第1章 はじめに

近年，ニューラルネットワークの発展に伴い，機械翻訳の翻訳精度が向上し，多くの形式的な文書（論文，法律文書等)は正しく翻訳することが可能になっている．しかしながら，生活や宗教など，文化的に独自の色合いの強い文章は，文化，思考と習慣に基づいて形成されているため，その言語にしか存在しない単語が含まれる場合がある．例えば，中国語で「汤圆」とは，中国伝統的なお祭りである「元宵節」に食べるもので，もち米の粉などで作られたゴマの入った球状で，主に似て食べるものという意味である．日本語には形で似たような食べ物はあるが，「元宵節」というお祭りで食するという文脈を表す単語が存在しない．そのため，「正月十五，家家户户吃汤圆，挂灯笼.」を機械翻訳で翻訳すると「朔望月の最初の 15 日目に，すべての家族が餃子を食べ，提灯を吊るします.」という翻訳結果が生成され，「汤圆」は「餃子」と誤訳されている．このような文章を正確に翻訳するには，文化固有単語の対訳の生成が必要不可欠である．

そこで，本研究では，異言語間の単語分散表現を用いて，原言語の文化固有の単語に対応する，目的言語の類似語を対訳として生成する．具体的には，日中それぞれの Wikipedia コーパスから各言語の分散表現空間を生成し，対訳辞書データに基づいて，両空間のアライメントを行う．次に，中国語の固有単語を日本語のベクトル空間に写像し，日本語の類義語を探すことで，固有単語の対訳を生成する．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．

類義語の組み合わせ

日中の分散表現空間の対応付けを行ったとしても，中国語の単語ベクトルと極めて近似する単語ベクトルを日本語の分散表現空間で得ることは難しい．対応する複数の日本語の類義語の中から中国語の単語を説明するのに有用な類義語の全組み合わせで新しいペアを生成して説明する必要がある．

類義語と差分の組み合わせ

中国語の単語で説明するのに類義語の全組み合わせで取得したペアでは，あまり意味を表現できない可能性がある．そのため，類義語ベクトルと中国語の単語ベクトルを別々に引き算で取得した最も近い日本語の単語と，それに対応する日本語の類義語を組み合わせ，新たに得られたペアを用いて中国語の単語を再度解釈する．

以下，本研究では２章において異文化から生じる文化的固有現象と，そこから生じる文化固有単語を翻訳する際の問題点を説明する．続いて，３章において単語分散表現を用いた文化固有単語の類似概念の選択方法を説明し，４章において，生成された文化固有単語の類似概念のペアの妥当性を評価する．そして，５章では 4 章の評価より，3 章で選定したペアの失敗例の原因と改善策を検討する．

第2章 文化固有単語

本章では，本研究で取り扱う文化固有単語について具体例を用いて説明する．また，文化固有単語の対訳生成の関連研究について記述する．

2.1 多言語コミュニケーションにおける問題

近年，機械翻訳を利用した多言語異文化コミュニケーションが増えている．例えば，ユーチューブの自動翻訳機能による字幕．また，コロナ禍で各国を自由に行き来できないため，オンラインで国境を越えたコラボレーションや交流などが盛んに行われている．しかしながら，異文化コミュニケーションでは，文化的背景や社会習慣，宗教観などが異なることが多く，それぞれの国に特有の文化現象は，その国の文化では同様の表現を見出すことができず，理解することが困難となる．このような状況で生じる単語は文化固有単語である．文化固有単語が運ぶ文化情報は，他の言語には「対応するもの」「同等なもの」がないため，対応する表現や直接的な解釈がない．そのため，機械翻訳で自動生成するとき，様々な誤訳や翻訳不可能な状況が発生する可能性がある．図 1 は機械翻訳の誤訳の具体例を示すものである．

例えば，中国の伝統的なお祭りである元宵節では，「汤圆」を食べる習慣がある．「汤圆」が伝統的な祭り，再会を祝うという意味に基づいているため，中国文化の下で文化的に特異な言葉に属し，他の国にはない文化現象である．

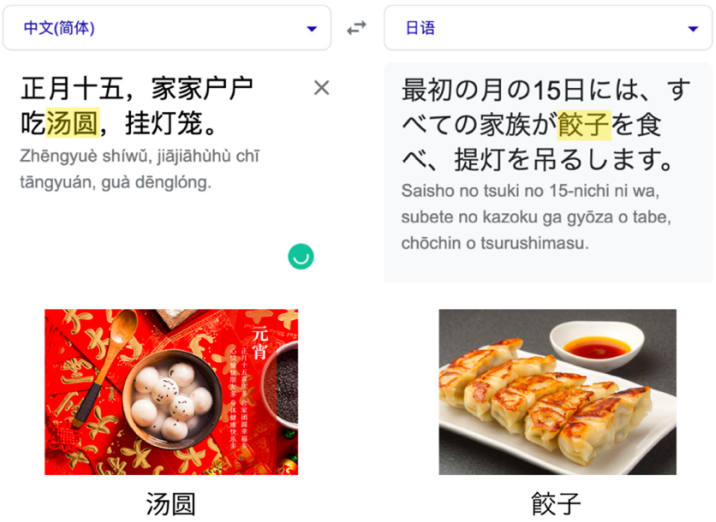

図 1：文化固有単語の誤訳の具体例

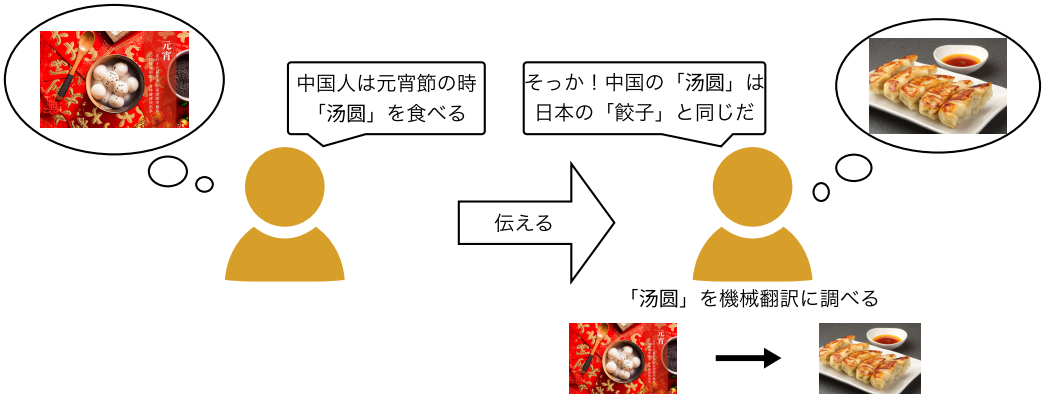

図 2: 文化固有単語の誤訳によるコミュニケーション齟齬の例

そのため,「汤圆」を含む文章を機械翻訳で訳すと,「汤圆」の文化的な意味が理解されず，誤って「餃子」と訳してしまう場合がある.

しかし,「汤圆」の意味を外国の方に伝える場合，このような誤解があると，話し手の伝える意味を理解できず，さらに大きな誤解を招くことになる．図２は外国の方が誤解する場合を説明する図である.

本研究では，このような異文化コミュニケーションする時に特有な文化現象で生じる言葉を文化固有単語とし，この文化固有単語の解釈を生成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固有単語の解釈に関する既存の研究を示す．文化固有単語は定義された固有の名詞ではないので，本研究は，未知語や異言語間での翻訳に関する研究を参考した.

未知語を探すため，田中らは単語の分散表現を用いて文書を分類する研究を行った．この研究は，医学に登場する様々な未知の単語に対し，単語表現空間を用いた文書を分類する[5]．2013 年にグーグルの研究者が公開された Word2Vec[1]を用いて日中それぞれの単語分散表現空間を作成できる．作成できた日中それぞれの単語分散表現空間を Alexis らが研究した CLWE(Cross-Lingual Word Embeddings)[4]に応用し，日中での単語分散表現空間を作ることができる．異言語間での単語分散表現空間の中に，藤川らは単語の意味を Word2Vec[1]によ

る分散表現を用いて表し，「参照訳を用いずに翻訳対象文と複数の訳候補文で異言語文間類似度を計算することで意味の観点から正しい訳を選択することができる［2][3］.」

第3章 文化固有単語の類似概念の検索

本章では，文化固有単語の類似概念を検索するための本研究でのアプローチを説明する．文化固有単語の類似概念を検索するために本研究では単語分散表現を用いて文化固有単語の類義語を選択する手法を提案する．

具体的にはまず，テキストデータをもとに日中の単語分散表現空間を作成する．言語によって文法が異なるため，同じ意味であっても，そこに含まれる単語のペアの位置が同じでない場合がある．ベクトル空間では，単語の位置にズレが生じると，その単語分散表現も変化する．このズレを補正するために異言語間単語埋め込みを行う．次に，対訳辞書で同一概念に紐付けされている日中それぞれの単語のベクトルを作成した単語分散表現空間から取り出す．そして，中国語固有単語ベクトルの周辺に存在する類義語 20 個を取得する．次に，取得した単語から，各単語の組み合わせについて単語ベクトルの平均を算出する．その平均ベクトルと中国語単語ベクトルとのコサイン類似度を算出することで，中国語固有単語を最もよく表現する組み合わせを探す．以下，それぞれのプロセスについて詳細を記述していく．

3.1 異言語間の分散表現空間の生成

本手法では Word2Vec を用いて単語分散表現空間を作成する．単語散布表現とは，文字や単語で埋め尽くされたベクトル空間上に，その文字や単語を表すベクトルである．Word2vec とは，単語をベクトルに変換し，単語のベクトルによって意味情報を表現する技術である．類似した単語はベクトル空間での座標が近く，無関係な単語は遠く離れていることがわかる．図 3 は「汤圆」を例として異言語間の分散表現空間での位置を表すものである．本手法では大量である Wikipedia の文章を使い，また，日中の二言語を用いて実験を行なった．単語分散表現空間の次元数は 300 とした．

3.1.1 単語分散表現空間の構築

本手法では Word2Vec を用いてそれぞれ日中の単語分散表現を作成する．単語分散表現とは，文字・単語をベクトル空間に埋め込み，その空間上のひとつの点として捉えることを目指し．単語埋め込み（Word Embedding）とも呼ばれる．

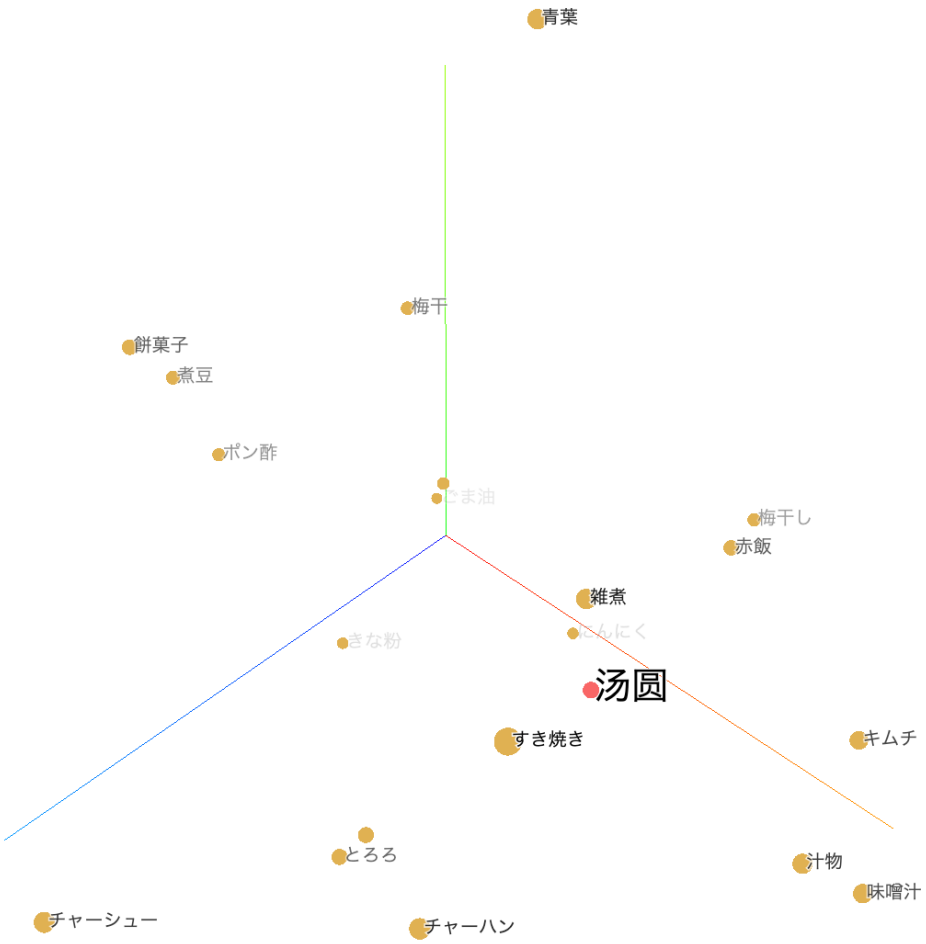

図 3：「汤圆」と日本語の類義語が分散表現空間での場所

Word2Vec は，文章中の単語を数値ベクトルに変換してその意味を把握する自然言語処理の手法である．ここでは，word2vec の skip-gram 法を用いて日中の単語分散表現を作る．skip-gram 法は，学習で中心のある単語から周辺の単語を予測する手法である．具体的には，教師あり学習で入力として中心語を与え，その周辺語の予測を出力する．この学習を通じて，ネットワークにある単語の周囲に．どのような単語が現れる可能性が高いのかを学習させる．

図 4 は skip-gram 法の説明を表したものである．本研究では，日中それぞれの Wikipedia コーパスから大量のデータを形態素解析で単語に分割してから学習することで各言語の分散表現空間を生成する．

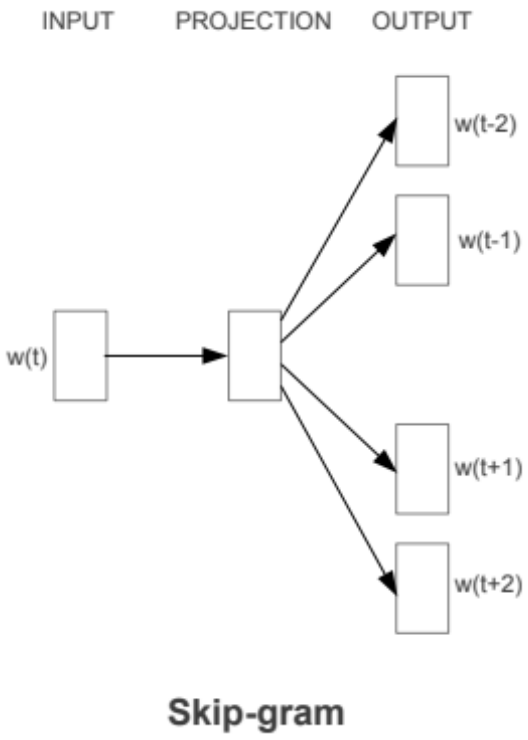

図４：skip-gram 法の説明[1]

3.1.2 単語分散表現空間の異言語間でのアライメント

異言語間単語埋め込みとは，異なる言語の単語を同じ単語ベクトル空間に埋め込むことで，同じ意味を持つが異なる言語の単語が同じベクトル表現になるようにすることである．本研究では Facebook が公開しているライブラリである MUSE を使用した．図 5 は異言語間単語埋め込みを行う時の図解である．

具体的には，まず，二つの単語分散表現空間を用意する．図５の（A）について，赤いのは英単語で，X で表示されている．青いのはイタリア語で， Y で示されている．そして，次の作業は，2 つの言語間の単語分散表現空間を整合・翻訳する．X と Y の各点には，その空間にある単語を表している． 点の大きさは，その言語の学習用コーパスに含まれる単語の頻度に比例している．次に，敵対的学習により，2 つの単語分散表現空間を大半一致させる回転行列 W を学習する．緑色の星印はランダムに選ばれた単語を識別器に供給され，2 つの単語分散表現空間が同じ分布にあるかどうかを判断する．

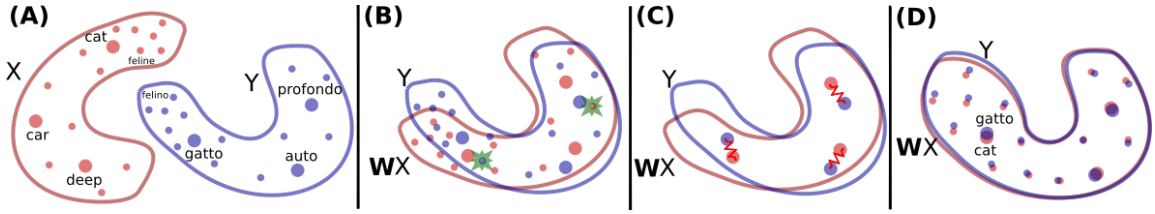

図５：異言語間単語埋め込みの図解[4]

このプロセスを図５の(B)で表している．また，プロクラステスによるマッピング W をさらに精緻化にさせる．この方法では，前ステップの頻出単語をアンカーポイントとして用い，対応するアンカーポイント間の距離を最小化する． その後，洗練されたマッピングを使用して，辞書に含まれるすべての単語をマッピングする．最後に，マッピング W と CSLS と呼ばれる距離尺度を用いて翻訳を行う． この尺度は，高密度な点が存在する空間を拡大した．このように，(A)の同じ区域と比べると，“中心”((A)の“cat”とか）は他の単語ベクトルとそれほど近くないになった．本研究では，図５で表した方法で，日中の単語分散表現空間をアライメントした．

まとめると，日本語の単語分散表現空間と中国語の単語分散表現を日本語の単語分散表現に近づけるように異言語単語埋め込みを行なった単語分散表現の二つを用いて本手法の評価を行なった．

具体的には 3.1.1 節で取得した日中それぞれ単語分散表現空間と日中対訳辞書を利用し，教師あり学習で中国語の単語ベクトル空間に近づけられた日本語の単語ベクトル空間を作成する．

3.2 文化固有単語の類義語の選択

日中の分散表現空間を対応させても，日本語の分散表現空間では，中国語の単語ベクトルに非常に近い単語ベクトルを得ることは困難である．そのため，この中国語に対応する有用な同義語を，対応するいくつかの日本語の類義語から選択する必要がある．

本手法では，本章で作成した単語分散表現空間内において，風習類 5 個と食品類 5 個，合計文化固有単語 10 個を検索する．また，これらの文化固有単語に対し，それぞれ 20 個の類似語を求めた．図 6 は日本語類義語を取得するアプローチを示すものである．そして，これらの類似語をそれぞれ組み合わせで全部 190 個のペアを作る．図 7 は日本語類義語ごとでのペアを作成するフローチャートを表すものである．また，文化固有単語と 20 個の日本語類義語がそれぞれ引

き算を行うことで，新したな日本語単語を取り出す．新した取得した日本語単語とそれらに対応する日本語類義語のペアを作成し，2 種類のペアを用いて文化固有単語を説明する．図 8 は差分により得られた日本語単語と日本語類義語のペアを作成するフローチャートを現したものである．

3.2.1 類義語の取得

本手法では，食品類 5 個と風習類 5 個，合計 10 個の文化固有単語を取り出すことになった．この 10 個の文化固有単語に関しては，以下のような特徴を持つことになる，①中国の文化背景の下でしか存在しない言葉であること．②選択された単語は，3.1.2 節で作成した単語分散表現空間に存在すること．この 2 つの条件をもとに，表 1 は文化固有単語の 10 件を現したものであり，表 2 は「汤圆」を具体例とし 20 件の日本語類義語の検索結果を現したものである．

表 1:文化固有単語 10 件

食品類	汤圆，肉夹馍，酒酿，凉皮，煎饼果子
風習類	炕，坐月子，交杯酒，数九，娃娃亲

表 2：「汤圆」の 20 件日本語類義語の検索結果

文化固有単語：汤圆			
類義語	コサイン類似値	類義語	コサイン類似値
①赤飯	0.6801488399505615	⑪チャーハン	0.6361185908317566
②餅菓子	0.6594136953353882	⑫ごま油	0.6343085765838623
③梅干し	0.6559799313545227	⑬キムチ	0.6310972571372986
④すき焼き	0.6479145884513855	⑭味噌汁	0.6283373236656189
⑤梅干	0.6450839638710022	⑮にんにく	0.6276915669441223
⑥枝豆	0.6427115201950073	⑯チャーシュー	0.6274588108062744
⑦青菜	0.6394413113594055	⑰ポン酢	0.6272783875465393
⑧きな粉	0.6388877630233765	⑱雑煮	0.626930296421051
⑨汁物	0.6381092667579651	⑲油揚げ	0.6264910101890564
⑩煮豆	0.6362212300300598	⑳とろろ	0.6245908141136169

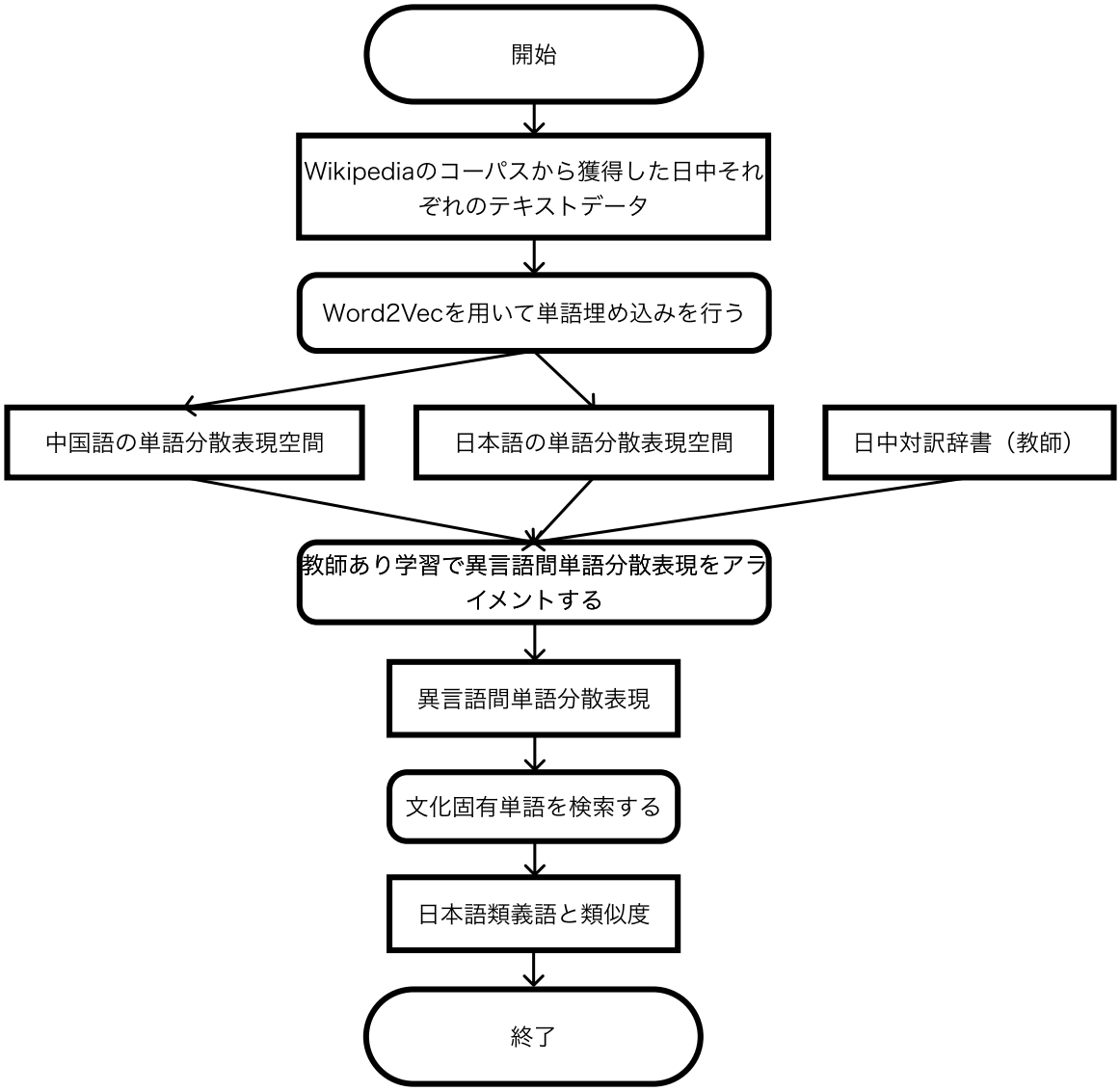

図 6：日本語類義語を取得するアプローチ

3.2.2 類義語の組み合わせ

3.2.1 節で得られ類義語だけでは，文化固有単語の対応する概念が日本語空間にあるかどうかを判断することは困難である．そこで，得られた 20 個の類義語を用いてさらに文化固有単語の対応する概念を絞り込む． 具体的には，まず得られた20個の類義語を全部組み合わせることにより,190 個のペアを作り出す．そして，作成したペアごとに足し算を行い，新しい単語ベクトルを算出する．また，算出した単語ベクトルと文化固有単語ベクトルのコサイン類似度をペアのランク値とし，適切性を評価する．ペアを降順でランキングすることで，最も類似した単語の組み合わせを見つけ選択する．表 3 は「汤圆」を例とし類義語の組み合わせの具体例を現したものである．

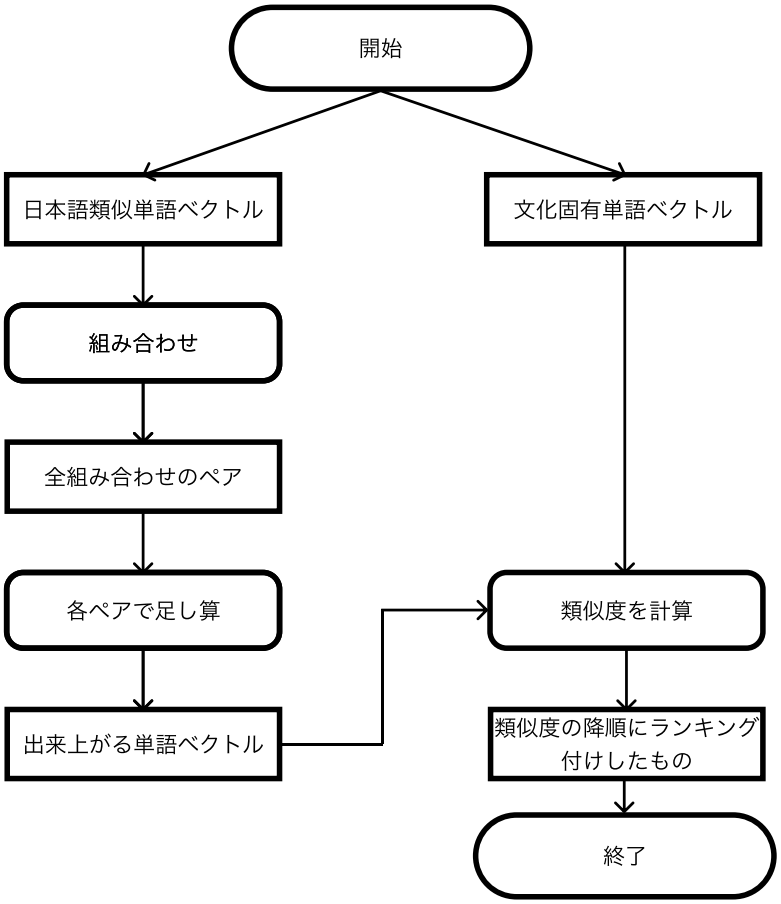

図 7：類義語の組み合わせ

表 3：「汤圆」の類義語の組み合わせ

文化固有単語：汤圆	
類義語のペア	ランク値
①(＇赤飯＇，＇すき焼き＇)	0.7173287868499756
②(＇赤飯＇，＇ごま油＇)	0.7170736789703369
③(＇すき焼き＇，＇ごま油＇)	0.7170068621635437
④(＇赤飯＇，＇チャーハン＇)	0.7154288291931152
⑤(＇赤飯＇，＇餅菓子＇)	0.7148652076721191
以下省略...	

3.2.3 類義語と差分の組み合わせ

具体的には，まず，対応させた 20 個の類義語ベクトルにおき，それぞれ文化固有単語ベクトルと引き算することで，差分ベクトルを算出する．取得した差分ベクトルにより，単語分散表現の中に差分ベクトルに最も近い日本語単語を抽出する．抽出された日本語単語とそれらに対応されている日本語の類義語を組み合わせでペアと作成する．ペアの適切性を確認するため，各ペアに足し算で取得したベクトルを用いて文化固有単語ベクトルとのコサイン類似度を算出する．算出されたコサイン類似度を類義語と差分のペアのランク値とし，適切性を評価する．ペアを降順でランキングすることで，最も類似した単語の組み合わせを発見する．表 4 は「汤圆」を例とし類義語と差分の組み合わせの具体例を現したものである．表 4 で挙げた例「汤圆」を見ると，ランク値 1 位は「＇餅菓子＇，＇赤飯＇」のペアになっていた．このペアの「赤飯」は文化固有単語の「汤圆」と日本語類義語 2 位の「餅菓子」の差分により得られた日本語単語である．このプロセスを図 9 に対応させると，まず「汤圆」①と「餅菓子」②を引き算することで，差分ベクトル③を取得する．そして，得られた差分ベクトル③に最も近い日本語単語「赤飯」④を抽出した．このように，「餅菓子」③と「赤飯」④を文化固有単語「汤圆」①の説明用のペアとする．次に，「餅菓子」③と「赤飯」④のペアに足し算することで和ベクトルを算出する．最後に，この和ベクトルと「汤圆」①のベクトルのコサイン類似度を計算し，「＇餅菓子＇，＇赤飯＇」のランク値とするものである．表 4 で挙げた数値のように，「＇餅菓子＇，＇赤飯＇」のランク値は 0.7148652076721191 であり

表 4：「汤圆」の類義語と差分の組み合わせ

文化固有単語：汤圆	
類義語と差分のペア	ランク値
①(＇餅菓子＇，＇赤飯＇)	0.7148652076721191
②(＇青菜＇，＇赤飯＇)	0.7117637991905212
③(＇梅干＇，＇月餅＇)	0.711032509803772
④(＇煮豆＇，＇赤飯＇)	0.708786129951477
⑤(＇ポン酢＇，＇月餅＇)	0.7078962922096252
以下省略...	

類義語の組み合わせで得られた「＇餅菓子＇，＇赤飯＇」のランク値が一致であることを確認できた．図 9 は上記の類義語と差分の組み合わせの流れを記したものである．

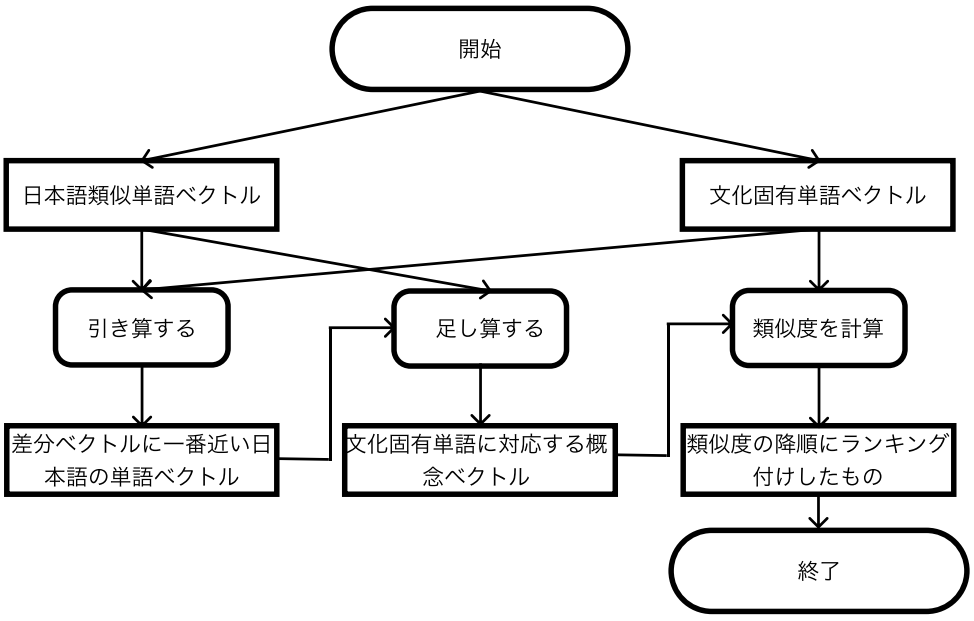

図 8：類義語と差分の組み合わせ

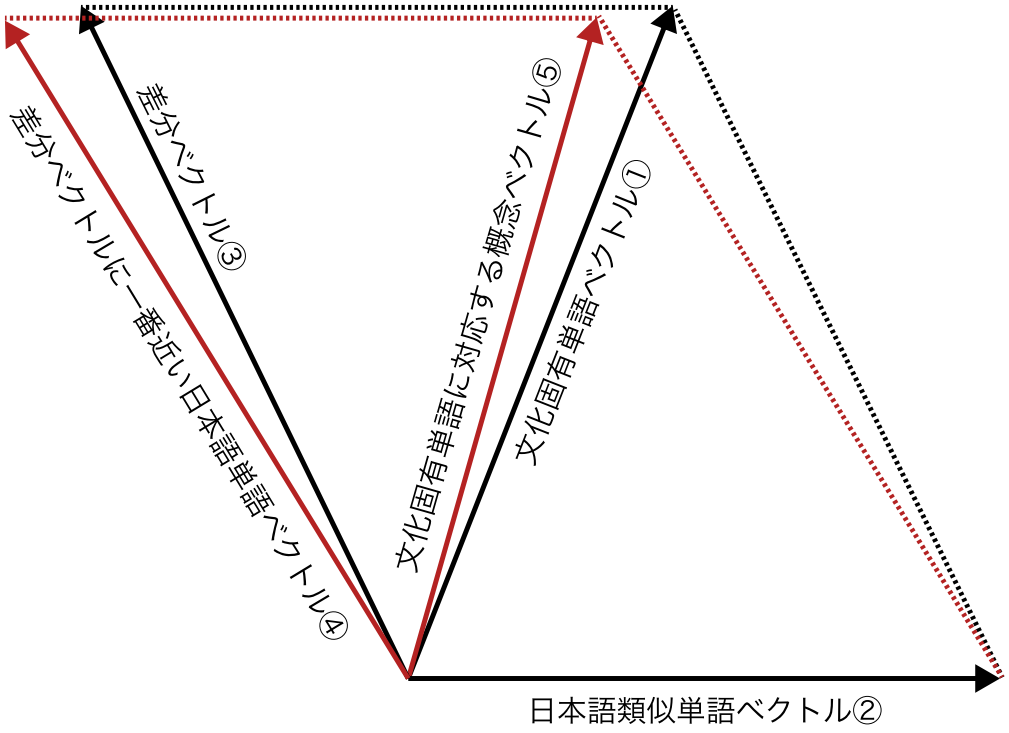

図９：類義語と差分の組み合わせの流れ

第4章 評価

4.1 評価指標

本章では 3 章の手法で取得した各ペア組の類似度やランク値を用いて評価する．逆ランクとは，検索結果から最初に正解が出る順位の逆数だという．得られた 20 個の日本語類義語と提案した 2 手法で取得したペアから人手で正解や不正解を判定し，平均逆ランクや平均適合率の平均を計算する．計算結果は以下の通りになった．表 5 が 3 種類の手法で算出された逆ランクや平均適合率を示し，図 10 が算出された逆ランクや平均適合率を表す箱図である．表 6 は取得した結果を用いて算出された平均逆ランクや平均適合率の平均を表しているものである．

表 5：逆ランクと平均適合率

文化固有単語		20 個の日本語類義語		類義語の組み合わせ		類義語と差分の組み合わせ	
		逆ランク	平均適合率	逆ランク	平均適合率	逆ランク	平均適合率
食品類	汤圆	1/2	0.5949	1/5	0.3229	1	0.5500
	肉夹馍	1/6	0.1937	1/5	0.2560	1/12	0.1186
	酒酿	1/15	0.0667	1/6	0.1667	1/15	0.0667
	凉皮	1/3	0.4381	1/2	0.5000	1	0.5635
	煎饼果子	1/17	0.0588	1/20	0.0500	1/12	0.1250
風習類	炕	1/3	0.2342	1/4	0.1513	1	0.6852
	坐月子	1	0.6337	1	0.5472	1	0.5064
	交杯酒	1/3	0.3005	1	0.6264	1	0.4894
	数九	1/4	0.3218	1/5	0.3389	1/5	0.3790
	娃娃亲	1/7	0.1429	1/17	0.0544	1/17	0.0544

表 6：平均逆ランクと平均適合率の平均

	平均逆ランク（MRR）	平均適合率の平均
20 個の日本語類義語	0.3185	0.2985
類義語の組み合わせ	0.3592	0.3014
類義語と差分の組み合わせ	0.5492	0.3538

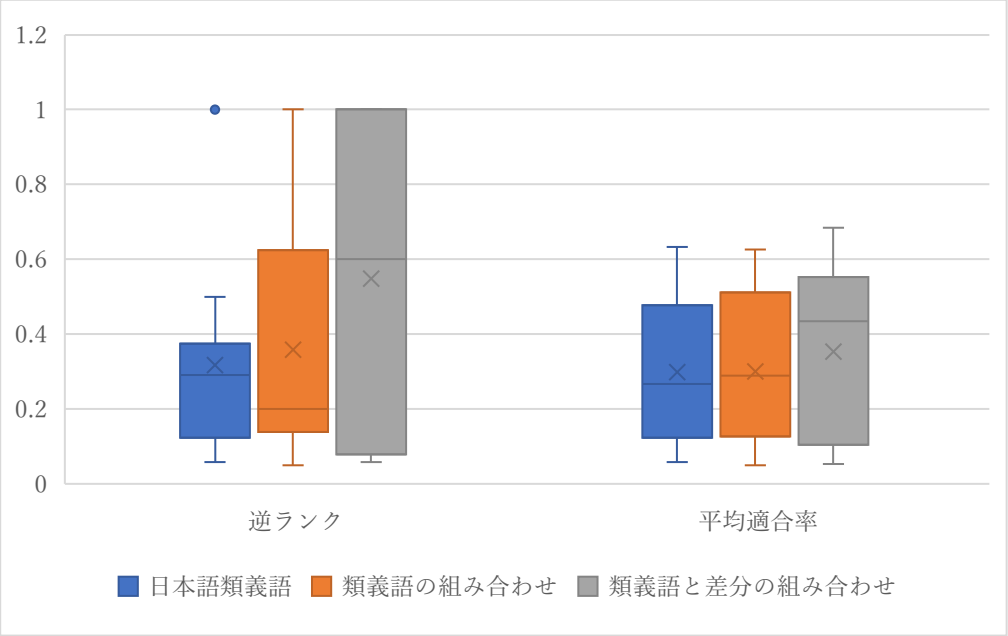

図 10：逆ランクと平均適合率の箱図

4.2 評価結果

本章では，平均逆ランクや平均適合率の平均を用いて中国語の単語をきちんと正しく説明されているかどうかを評価する．4.1 節に導き出した表 6 の通り，10 個の文化固有単語に対し，日本語の類義語の平均逆ランクは 0.3185 であり，平均適合率の平均は 0.2985 であることを示している．類義語の組み合わせで，平均逆ランクは 0.3592 になり，単純の類義語の説明より平均逆ランク(MRR)の上昇率が 12.78%であり，平均適合率の平均の上昇率が 97.15%であることがわかる．また，類義語と差分の組み合わせで出した結果と日本語の類義語から得られた結果を比較すると，平均逆ランク(MRR)が 72.43%に向上し，平均適合率の平均のが 18.53%に向上することが成功した．

しかし，表 5 の結果から見ると，文化固有単語の類似概念の検索結果が全部成功したわけではない．表 5 で出した逆ランクから，提案手法を使用した後食品類の「凉皮」が最初に正解になる順位が改善されたので，成功例と判定された．逆に，単純に日本語の類義語と比べ，類義語と差分の組み合わせで「肉夹馍」が最初に正解が出る順位が 6 位に下がり，平均適合率も 38.77%に低下した．同様に，風習類で明確に成功と判断できる例として「交杯酒」がある．なぜなら，類義語の組み合わせや類義語と差分の組み合わせに最初に正解が出る順位が同じく 1

位になり，平均適合率もそれぞれ 2 倍と 62.86%上昇したからである．しかし，失敗例は「娃娃亲」とわかる．提案手法の使用後，「娃娃亲」が最初に正解が出る順位がそれぞれ 17 位に下がり，平均適合率も両者で 61.93%低下したためである．

第5章 考察

本章では，文化固有単語の類似概念の生成の成功例と失敗例について分析し，改善案を考える．

成功例

表７は成功例を示したものである．課題として，成功と認めている中国語の単語には以下の特徴がある．

まず，「汤圆」と「交杯酒」は特定の場合に存在するものである．例えば「汤圆」には基本「元宵節」（お祭り）に食べるものだと認識され，つまり「お祭り」と「食べ物」という二つの意味を含めている．本研究で使用した三つの検索方法から取得した結果におき，まず，日本語の類義語で「赤飯」と「餅菓子」が現れ，つまり「お祭り」と「食べ物」に関連する類似度の高い単語を登場した．次に，類義語と差分の組み合わせを用いて類似概念のペアを探すとき，「汤圆」と「餅菓子」の差分から最も近い日本語の単語に「赤飯」という単語を成功に見つかった．それだけではなく，類義語の組み合わせのペアでも「赤飯」＋「餅菓子」というペアを発見した．また，「赤飯」＋「餅菓子」のペアでは，どちらの提案手法でも同じ順位値を達成することがわかった．これは，2 つの提案手法を用いることで，「汤圆」の「お祭りの時に食べるもの」という意味を説明できる組み合わせを見つけることに成功したことを示している．このように，単純の日本語の類義語の結果と比べると，日本語の類義語で「赤飯」＋「餅菓子」という正解の順位が後に登場することにかかわらず，「汤圆」はまた成功例だと考えることができる．

逆に，「交杯酒」には「結婚式での新郎新婦が腕を組んで一緒に飲むお酒」であり，同時に「結婚式」＋「新郎新婦」＋「腕組み」＋「お酒」の四つの意味を含めている．「汤圆」と比較すると，一つの単語にさらに二つの意味を含めているせいかもしれないので，「汤圆」のような良い結果をなっていない．それにもかかわらず，提案手法の使用後は正解率が向上していることが確認できた．

そして，失敗例と比べ，成功例には一つの単語になっている．つまり「単語 1」＋「単語 2」の形式ではないので，最初に中国語の単語分散表現空間を作る時に，適所に現れる可能性が高くなった．したがって，このような単語に提案手法を狙う方が有効になっているかを考えている．

表 7：提案手法による正解の生成例

		最初に正解が出た順位		
成功例		日本語類義語	類義語の組み合わせ	類義語と差分の組み合わせ
食品類	汤圆	2	5	1（1↑）
風習類	交杯酒	3	1（2↑）	1（2↑）

失敗例

表 8 は失敗例を示したものである．成功例と異なり，失敗した中国語の単語自身には複数の意味が含めている．表 8 で挙げられる単語の「肉夹馍」や「娃娃亲」のどちらにも，「単語１」＋「単語２」の組み合わせで生成した単語である．よって．中国語の単語分散表現空間を作る時に，形態素解析により，文を分割して単語を取得するとき，誤りが起こる可能性を考えている．従って，このような単語は中国語の単語分散表現空間の正しい場所にいない可能性も高いし，そのまま日中の単語分散表現空間をアライメントすると，対応に間違いがあるのかを考えている．中国語の単語そのものの意味を誤解する可能性があるため，異言語間での単語分散表現空間では正しい解釈ができない．

他にも，教師あり学習で異言語間での単語分散表現空間を作成したが，今回教師として使用した日中対訳辞書は，グーグル翻訳により作成したものである．機械翻訳を用いて生成した対訳辞書であるため，精度が保証できないので，日中それぞれ単語分散表現空間を回転対応する際の精度があまり高くないかもしれない．

以上の二つの原因を検討する上で改善案を提出する．原因１について，提案手法だけではなく，複数の対訳テンプレート文（「XX（関連する中国語単語）でする（動詞）XX（中国語単語）」）を定義し，中国語の文化固有単語だけでなく，その単語に関連する単語を用いてフレーズを生成する．そのフレーズの生成に用いたテンプレートを利用し，日本語の複数の類似概念を当てはめることで，日本語の説明文を生成する．つまり，一つの単語から説明ペアを探すだけではなく，単語自身にさらに適切な意味を絞り込むことで最も良い説明を発見したい．原因２について，日中対訳辞書には，専門的な電子版の対訳辞書を使用することにより，単語分散表現空間の精度を向上させることで，より類似度の高い類似概念を得ることができる．

表 8：提案手法による誤訳の生成例

		最初に正解が出た順位		
失敗例		日本語類義語	類義語の組み合わせ	類義語と差分の組み合わせ
食品類	肉夾饃	6	12（10↓）	5
風習類	娃娃亲	7	17（10↓）	17（10↓）

第6章 おわりに

多言語コミュニケーションにおける文化固有単語の説明を生成するために，本研究では単語分散表現から得られるベクトルを多言語の単語分散表現空間から取得し，選択･合成するアプローチを提案してきた．本研究の貢献は以下の通りである．

類義語の組み合わせ

アライメントを行った日中の分散表現空間を用いて，中国語の文化固有単語に類似する日本語の単語 20 件を取得し，全単語ペアの平均ベクトルの類似度よって，順序付けを行った．降順でランキングしたものから中国語の単語を適切に解釈できる組み合わせの数を決定した．また，単純に日本語の類義語の大役と類義語の組み合わせの平均逆ランクや平均適合率の平均を比較する上で，平均逆ランク(MRR)が 4.07%，平均適合率の平均（MAP）が約 0.29%向上したため，本手法は有効であることを確認した．

類義語と差分の組み合わせ

類義語の組み合わせだけでなく，類義語と差分の組み合わせも用いることで，中国語の文化固有単語の説明を生成した．提案手法により，単純に日本語の類似単語を対訳とする場合と比べて，平均逆ランク(MRR)が 23.07%，平均適合率の平均（MAP）が 5.53%向上した．本研究の提案手法が中国語の単語の解釈が有効であることがわかる．

異言語間での単語分散表現空間から，類義語と類義語の組み合わせと類義語と差分の組み合わせを抽出して用いることで中国語の単語を解釈することができた．だが，得られた結果では，やはり正解の順位が低下し，平均適合率が低下した場合がある．本手法を異文化コミュニケーションの場で使用するには，正解の順位や平均適合率をさらに高める必要がある．

謝辞

本研究を行うあたり，ご指導していただいた村上陽平准教授に深く感謝申し上げます．また，支えてくださった同研究室の大井也史先輩に深く感謝申し上げます．

参考文献

[1] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean : Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[2] 藤川寛基, 越前谷博, 荒木健治 : 単語の分散表現を用いた異言語文間類似度に基づく最適訳選択, 第 16 回情報科学技術フォーラム(FIT2017)講演論文集, 第2分冊, pp.185-186, (2017) .

[3] 藤川寛基, 越前谷博, 荒木健治 : 参照訳を必要としない単語分散表現による異言語間類似度を用いた訳文の自動評価[J]. 研究報告知能システム (ICS), 2018(10): pp.1-6. (2018)

[4] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, Herve Jegou : Word translation without parallel data, arXiv preprint arXiv:1710.04087 (2017).

[5] 田中昌昭 : 単語の分散表現を用いた文書分類[J]. 川崎医療福祉学会誌, 28(1): pp.167-178. (2018)

付録：図

A.1 食品類

「汤圆」

日本語類似単語	正解・不正解	類義語と差分ベクトルに一番近い日本語のペア	正解・不正解	日本語類義語のペア	正解・不正解
('赤飯', 0.6801488399505615)		(('餅菓子', '赤飯'), 0.7148652076721191)	正解	(('赤飯', 'すき焼き'), 0.7173287868499756)	
('餅菓子', 0.6594136953353882)	正解	(('青菜', '赤飯'), 0.7117637991905212)		(('赤飯', 'ごま油'), 0.7170736789703369)	
('梅干し', 0.6559799313545227)		(('梅干', '月餅'), 0.711032509803772)		(('すき焼き', 'ごま油'), 0.7170068621635437)	
('すき焼き', 0.6479145884513855)		(('煮豆', '赤飯'), 0.708786129951477)		(('赤飯', 'チャーハン'), 0.7154288291931152)	
('梅干', 0.6450839638710022)		(('ポン酢', '月餅'), 0.7078962922096252)		(('赤飯', '餅菓子'), 0.7148652076721191)	正解
('枝豆', 0.6427115201950073)		(('ごま油', '桃の節句'), 0.7077682018280029)		(('赤飯', 'にんにく'), 0.7144782543182373)	
('青菜', 0.6394413113594055)		(('赤飯', 'ラヴィーダ'), 0.6934508681297302)		(('赤飯', 'チャーシュー'), 0.7137669324874878)	
('きな粉', 0.6388877630233765)	正解	(('枝豆', '餡'), 0.682388424873352)		(('赤飯', '枝豆'), 0.7132213115692139)	
('汁物', 0.6381092667579651)	正解	(('梅干し', 'シェンディ・グループ'), 0.6737633347511292)		(('赤飯', 'キムチ'), 0.7130994200706482)	
('煮豆', 0.6362212300300598)		(('すき焼き', 'ヨンヒョン'), 0.662400484085083)		(('すき焼き', 'きな粉'), 0.7122315168380737)	正解
('チャーハン', 0.6361185908317566)		(('チャーハン', ']]) [[', 0.6545721292495728)		(('赤飯', '青菜'), 0.7117637991905212)	
('ごま油', 0.6343085765838623)		(('チャーシュー', '祓除'), 0.6544845700263977)		(('餅菓子', '梅干'), 0.7108181118965149)	正解
('キムチ', 0.6310972571372986)		(('汁物', 'ゴールデンハニー'), 0.654338002204895)	正解	(('赤飯', '梅干'), 0.7102778553962708)	
('味噌汁', 0.6283373236656189)		(('油揚げ', 'マンスリーアシスタント'), 0.6504554748535156)		(('赤飯', 'きな粉'), 0.7100368738174438)	正解
('にんにく', 0.6276915669441223)		(('にんにく', '三つ指'), 0.6498146057128906)		(('梅干し', 'すき焼き'), 0.7096458673477173)	
('チャーシュー', 0.6274588108062744)		(('キムチ', '鏃型'), 0.6483259201049805)		(('赤飯', '煮豆'), 0.708786129951477)	
('ポン酢', 0.6272783875465393)		(('味噌汁', 'フッフティング'), 0.6395987272262573)		(('赤飯', '梅干し'), 0.7067073583602905)	
('雑煮', 0.626930296421051)	正解	(('雑煮', 'クリッカ'), 0.6387474536895752)	正解	(('すき焼き', 'にんにく'), 0.7065438032150269)	
('油揚げ', 0.6264910101890564)		(('とろろ', 'カンパネラ'), 0.6252217888832092)		(('餅菓子', '煮豆'), 0.7037994861602783)	
('とろろ', 0.6245908141136169)		(('きな粉', '祭り'), 0.5154975652694702)	正解	(('餅菓子', 'ごま油'), 0.7033324241638184)	

「肉夹馍」

日本語類似単語	正解・不正解	類義語と差分ベクトルに一番近い日本語のペア	正解・不正解	日本語類義語のペア	正解・不正解
('ほうじ茶', 0.4799642860889435)		(('煮豆', 'エントランス'), 0.5181539058685303)		(('ほうじ茶', '奈良漬け'), 0.5281751155853271)	
('奈良漬け', 0.47709375619888306)		(('奈良漬', 'ドコモサポートデスク'), 0.5133342146873474)		(('ほうじ茶', 'もんじゃ焼き'), 0.5281003713607788)	
('煮豆', 0.4715301990509033)		(('テンジャン', 'ライコランド'), 0.5018234848976135)		(('ほうじ茶', '奈良漬'), 0.5275562405586243)	
('漬物', 0.46796461939811707)		(('ほうじ茶', 'ユーゴスラビア・ドルガ・リーガ'), 0.4979695677757263)		(('奈良漬け', 'わさび漬け'), 0.5275432647460632)	
('奈良漬', 0.46006715297698975)		(('ナンプラー', 'ストロイエ'), 0.4885004460811615)		(('ほうじ茶', '包子'), 0.5204176902770996)	正解
('麺類', 0.45564234256744385)	正解	(('蒸し物', '焚聞'), 0.47976261377334595)		(('ほうじ茶', '餅菓子'), 0.5176457762718201)	
('餅菓子', 0.447515070438385)		(('漬物', 'クロエドフ'), 0.4763653576374054)		(('奈良漬け', '奈良漬'), 0.5162889957427979)	
('鍋物', 0.4465087354183197)		(('餅菓子', 'シュテティン'), 0.47398626804351807)		(('奈良漬', 'もんじゃ焼き'), 0.5161566734313965)	
('菓子パン', 0.4436976909637451)		(('もんじゃ焼き', '焚聞'), 0.47196999192237854)		(('ほうじ茶', '竹輪'), 0.5158421397209167)	
('チャーシュー', 0.4428340792655945)	正解	(('竹輪', 'キルクス・マキシムス'), 0.4688947796821594)		(('煮豆', '奈良漬'), 0.5146343111991882)	
('揚げ物', 0.4423879086971283)		(('鍋物', 'ニッポリヒト'), 0.46674713492393494)		(('奈良漬', 'わさび漬け'), 0.5137489438056946)	
('竹輪', 0.44182083010673523)		(('チャーシュー', '旬府'), 0.4661326408386230 5)	正解	(('ほうじ茶', 'わさび漬け'), 0.5124591588973999)	
('わさび漬け', 0.4415242075920105)		(('包子', 'GROND'), 0.45927679538726807)	正解	(('奈良漬', '蒸し物'), 0.5115745067596436)	
('包子', 0.44008943438352997)	正解	(('麺類', 'ジェノシード'), 0.4587852656841278)	正解	(('ほうじ茶', 'ナンプラー'), 0.5114549994468689)	
('ナンプラー', 0.43778523802757263)		(('菓子パン', 'AERS'), 0.45600587129592896)		(('奈良漬け', '餅菓子'), 0.5111149549484253)	
('明太子', 0.43777996301651)		(('明太子', 'サンタ・マリーア'), 0.4536254107952118)		(('奈良漬け', 'もんじゃ焼き'), 0.5109967589378357)	
('テンジャン', 0.43687865138053894)		(('揚げ物', 'クロエドフ'), 0.4525320827960968)		(('奈良漬', '包子'), 0.5097033381462097)	正解
('蒸し物', 0.43596136569976807)		(('削り節', 'ラーフェンブスリュック'), 0.45250579714775085)		(('奈良漬け', '煮豆'), 0.5092515349388123)	
('削り節', 0.43530717492103577)		(('わさび漬け', '長江大橋'), 0.4272901713848114)		(('煮豆', 'もんじゃ焼き'), 0.508973240852356)	
('もんじゃ焼き', 0.43455538153648376)		(('奈良漬け', 'TSV'), 0.24174292385578156)		(('ほうじ茶', '煮豆'), 0.5089459419250488)	

「酒酿」

日本語類似単語	正解・不正解	類義語と差分ベクトルに一番近い日本語のペア	正解・不正解	日本語類義語のペア	正解・不正解
('にんにく', 0.6234809756278992)		(('山芋', 'ジェリエズニチャル'), 0.6388952136039734)		(('にんにく', '包子'), 0.6778224110603333)	
('煮豆', 0.6138449311256409)		(('煮豆', 'ローレンス・タルボット'), 0.6349256634712219)		(('にんにく', '筍'), 0.676318347454071)	
('緑豆', 0.6128088831901555)		(('にんにく', 'デスパライア'), 0.6294365525245667)		(('山芋', '包子'), 0.6747486311106873)	
('山芋', 0.6122388243675232)		(('山葵', 'フォーティセブン'), 0.6220980286598206)		(('包子', '玉露'), 0.6746395230293274)	
('杏仁', 0.5995081067085266)		(('メンマ', 'アメリカンサッカーリーグ'), 0.6197900176048279)		(('筍', 'ごま油'), 0.6728888750076294)	
('とろろ', 0.5979765057563782)		(('オイスターソース', '臨尼'), 0.61966866254806 52)		(('包子', '糠漬け'), 0.6707261800765991)	
('筍', 0.5966622233390808)		(('包子', 'チュスタ'), 0.6195994019508362)		(('杏仁', '玉露'), 0.6691166758537292)	
('青菜', 0.5959372520446777)		(('緑豆', 'デスパライア'), 0.6194961071014404)		(('とろろ', '玉露'), 0.6688677072525024)	
('ごま油', 0.5954062938690186)		(('杏仁', 'インジヤ'), 0.6130394339561462)		(('包子', '胡瓜'), 0.668691873550415)	
('メンマ', 0.595319926738739)		(('ほうじ茶', 'マルクス・オクタウィウス'), 0.6111831068992615)		(('にんにく', '山葵'), 0.6678009033203125)	
('こんにゃく', 0.5941697359085083)		(('ごま油', 'ローレンス・タルボット'), 0.6108680367469788)		(('緑豆', '山葵'), 0.6666301488876343)	
('包子', 0.5933231115341187)		(('青菜', 'デスパライア'), 0.6074937582015991)		(('煮豆', '包子'), 0.6663436889648438)	
('胡瓜', 0.5899393558502197)		(('とろろ', 'ルクワイア'), 0.6037170886993408)		(('包子', 'ほうじ茶'), 0.6656651496887207)	
('茄子', 0.5892964601516724)		(('筍', 'アッシュバッハー'), 0.6014594435691833)		(('煮豆', '山葵'), 0.6653916835784912)	
('糠漬け', 0.5890556573867798)	正解	(('糠漬け', 'デスパライア'), 0.6002573370933533)	正解	(('糠漬け', '山葵'), 0.6643429398536682)	正解
('玉露', 0.5889763832092285)		(('胡瓜', 'ダイレオン'), 0.599047839641571)		(('にんにく', '茄子'), 0.6639174818992615)	
('枝豆', 0.5886611342430115)		(('こんにゃく', 'デスパライア'), 0.5987100601196289)		(('にんにく', '杏仁'), 0.6633684635162354)	
('ほうじ茶', 0.5882912278175354)		(('玉露', 'デスパライア'), 0.5976647138595581)		(('緑豆', '玉露'), 0.6632915735244751)	
('山葵', 0.5870627164840698)		(('枝豆', 'デスパライア'), 0.5962415933609009)		(('ごま油', '包子'), 0.6631489396095276)	
('オイスターソース', 0.586053192615509)		(('茄子', 'エブスコ'), 0.5944491028785706)		(('山芋', '玉露'), 0.6625866889953613)	

「涼皮」

日本語類似単語	正解・不正解	類義語と差分ベクトルに一番近い日本語のペア	正解・不正解	日本語類義語のペア	正解・不正解
('煮豆', 0.5854130387306213)		(('ごま油', '3230'), 0.6310734748840332)	正解	(('明太子', 'タケノコ'), 0.6316536664962769)	
('山芋', 0.576465904712677)		(('煮豆', 'フッフティング'), 0.6184390187263489)		(('辛子', '胡瓜'), 0.6303043961524963)	正解
('ごま油', 0.5763699412345886)	正解	(('梅干', 'JUNKSTA'), 0.6042561531066895)		(('胡瓜', '餅菓子'), 0.6275327205657959)	
('辛子', 0.5758215188980103)	正解	(('山芋', 'ジェリエズニチャル'), 0.5936452150344849)		(('ごま油', '胡瓜'), 0.6257614493370056)	正解
('胡瓜', 0.5722225308418274)	正解	(('餅菓子', 'ブッシュヒュッテン'), 0.5899266004562378)		(('辛子', 'タケノコ'), 0.6254138946533203)	正解
('黒豆', 0.5712259411811829)		(('辛子', 'エイグロフ'), 0.5880473256111145)	正解	(('明太子', 'ゴボウ'), 0.6248966455459595)	
('梅干', 0.5701934099197388)		(('胡瓜', 'ダイレオン'), 0.5819050073623657)	正解	(('山芋', 'ほうじ茶'), 0.6248610019683838)	
('青菜', 0.5680230259895325)		(('ポン酢', '焚間'), 0.578225314617157)	正解	(('ごま油', '明太子'), 0.6244164705276489)	
('にんにく', 0.5659486651420593)	正解	(('にんにく', 'ユーゴスラビア・ドルガ・リーガ'), 0.5774382948875427)	正解	(('ほうじ茶', '餅菓子'), 0.6237940788269043)	
('ほうじ茶', 0.5658571124076843)		(('ほうじ茶', 'ユーゴスラビア・ドルガ・リーガ'), 0.5772449374198914)		(('ごま油', '辛子'), 0.6231295466423035)	正解
('きな粉', 0.559394896030426)		(('黒豆', 'デスパライア'), 0.576265275478363)		(('胡瓜', 'テンジャン'), 0.622790515422821)	
('漬物', 0.5555298924446106)		(('青菜', 'デスパライア'), 0.574464738368988)		(('辛子', '黒豆'), 0.6223971247673035)	
('餅菓子', 0.5550835728645325)		(('糠漬け', '˝⊕､'), 0.5729289650917053)		(('胡瓜', 'ほうじ茶'), 0.6215317249298096)	
('明太子', 0.554932713508606)		(('きな粉', 'デラップ'), 0.5694428086280823)		(('ごま油', '餅菓子'), 0.6214209794998169)	
('テンジャン', 0.554572582244873)		(('明太子', 'エイグロフ'), 0.5656202435493469)		(('胡瓜', '明太子'), 0.6213462352752686)	
('ポン酢', 0.5492446422576904)	正解	(('漬物', 'クロエドフ'), 0.561231791973114)		(('煮豆', '胡瓜'), 0.621262788772583)	
('ゴボウ', 0.5481053590774536)		(('黒砂糖', '5825'), 0.5590115189552307)		(('ごま油', 'ほうじ茶'), 0.6207605600357056)	
('タケノコ', 0.547924876213073 7)		(('タケノコ', 'デスパライア'), 0.5519115328788757)		(('山芋', '辛子'), 0.6206479072570801)	
('黒砂糖', 0.5459280014038086)		(('ゴボウ', 'デスパライア'), 0.5504387021064758)		(('山芋', '明太子'), 0.6203501224517822)	
('糠漬け', 0.5439153909683228)		(('テンジャン', '小刀'), 0.4773220419883728)		(('山芋', '胡瓜'), 0.6202166676521301)	

「煎餅果子」

日本語類似単語	正解・不正解	類義語と差分ベクトルに一番近い日本語のペア	正解・不正解	日本語類義語のペア	正解・不正解
('カルルッチョズ', 0.42296648025512695)		(('エボキシエタン', '水炊き'), 0.4707833230495453)		(('奈良漬け', 'エボキシエタン'), 0.5011557340621948)	
('マリネード', 0.4170953333377838)		(('スンニュン', 'Zagier'), 0.4601863920688629)		(('奈良漬け', '鎮下'), 0.4939696490764618)	
('奈良漬け', 0.4159676730632782)		(('フォワグラ', '鹿の子絞'), 0.4542381465435028)		(('エボキシエタン', 'アサムラサキ'), 0.490120530128479)	
('芥子菜', 0.4116165339946747)		(('奈良漬け', 'アントニオ・リュディガー'), 0.4524662494659424)		(('鎮下', 'フォワグラ'), 0.4890938699245453)	
('エボキシエタン', 0.41144803166389465)		(('ぬかづけ', '粉米'), 0.4445470869541168)		(('カルルッチョズ', 'フィッシュミール'), 0.4888852834701538)	
('鎮下', 0.4112981855869293)		(('芥子菜', '水炊き'), 0.44301465153694153)		(('カルルッチョズ', '鎮下'), 0.4872332811355591)	
('フィッシュミール', 0.40784209966659546)		(('水炊き', 'サンタ・マリーア'), 0.4159084260463714 6)		(('エボキシエタン', '鎮下'), 0.4872084856033325)	
('ナシルマッ', 0.40569785237312317)		(('大和芋', 'アメリカンスムース'), 0.4144882559776306)		(('奈良漬け', 'ケニー・ビー'), 0.48614102602005005)	
('ケニー・ビー', 0.4040181040763855)		(('カルルッチョズ', 'マテ'), 0.401893675327301)		(('鎮下', 'ナシルマッ'), 0.4860389828681946)	
('ぬかづけ', 0.40207603573799133)		(('ソトアヤム', 'マラソンイメージソング'), 0.40018099546432495)		(('エボキシエタン', '大和芋'), 0.48469531536102295)	
('チャプチェ', 0.3998231589794159)		(('ほうじ茶', 'ヴァルガード'), 0.3976046144962311)		(('カルルッチョズ', 'アサムラサキ'), 0.4838443100452423)	
('アサムラサキ', 0.39856454730033875)		(('ケニー・ビー', '辛子'), 0.3905361294746399)	正解	(('エボキシエタン', 'スンニュン'), 0.4826045334339142)	
('大和芋', 0.3978874087333679)		(('アサムラサキ', '粉食'), 0.3850493133068084 7)		(('ナシルマッ', 'アサムラサキ'), 0.4825762808329065)	
('デュラムコムギ', 0.3961288034915924)		(('鎮下', 'もやし'), 0.3832284212112427)		(('鎮下', 'アサムラサキ'), 0.4810434579849243)	
('水炊き', 0.3957955539226532)		(('デュラムコムギ', '盛り上がれ'), 0.3764277994632721)		(('カルルッチョズ', 'エボキシエタン'), 0.48098477721214294)	
('スンニュン', 0.3953886032104492)		(('choux', 'トミーズ'), 0.3549750149250030 5)	正解	(('ケニー・ビー', 'フォワグラ'), 0.4800265431404114)	
('choux', 0.39457574486732483)	正解	(('フィッシュミール', 'トミーズ'), 0.3498767614364624)		(('カルルッチョズ', 'デュラムコムギ'), 0.47994300723075867)	
('フォワグラ', 0.3939008414745331)		(('チャプチェ', 'ペい'), 0.3064014613628387 5)		(('鎮下', 'ソトアヤム'), 0.47982102632522583)	
('ほうじ茶', 0.3938712775707245)		(('マリネード', '前説'), 0.28005895018577576)		(('エボキシエタン', 'フォワグラ'), 0.479763001203537)	
('ソトアヤム', 0.3912789821624756)		(('ナシルマッ', '大林'), 0.25707370042800903)		(('鎮下', 'スンニュン'), 0.47911038994789124)	正解

A.2 風習類

「炕」

日本語類似単語	正解・不正解	類義語と差分ベクトルに一番近い日本語のペア	正解・不正解	日本語類義語のペア	正解・不正解
('囲炉裏', 0.5354833006858826)		(('明障子', '寝そべる'), 0.5944862961769104)	正解	(('土間', '注連縄'), 0.6018885374069214)	
('連子窓', 0.5341041684150696)		(('遣戸', '囲炉裏'), 0.5830212235450745)	正解	(('連子窓', 'しめ縄'), 0.5999647378921509)	
('土間', 0.5320331454277039)	正解	(('しめ縄', '土間'), 0.5813121795654297)		(('囲炉裏', '鏡板'), 0.597690999507904)	
('明障子', 0.5258412957191467)		(('板敷き', '注連縄'), 0.5626183152198792)		(('囲炉裏', '明障子'), 0.5966817736625671)	正解
('板敷き', 0.5110518336296082)		(('連子窓', '竈'), 0.5529605150222778)		(('板敷き', 'しめ縄'), 0.5950745940208435)	
('衝立', 0.5099860429763794)		(('囲炉裏', '6284'), 0.5502819418907166)		(('囲炉裏', '連子窓'), 0.5934620499610901)	正解
('桟', 0.5072023868560791)		(('板敷', '灯明'), 0.5432043075561523)		(('囲炉裏', '衝立'), 0.5919680595397949)	
('蔀', 0.5071710348129272)		(('土間', 'ヴィキングル'), 0.5367313027381897)	正解	(('明障子', 'しめ縄'), 0.5918456315994263)	
('注連縄', 0.5025330781936646)		(('蔀', '打ち込める'), 0.5355149507522583)		(('囲炉裏', '蔀'), 0.5910852551460266)	
('縁側', 0.5014920830726624)		(('軒下', '墾局'), 0.525785505771637)		(('囲炉裏', '桟'), 0.5908016562461853)	
('軒下', 0.49849170446395874)		(('注連縄', '6061'), 0.521374523639679)		(('囲炉裏', '板敷'), 0.5897334218025208)	
('障子', 0.4966762661933899)		(('桟', 'ふれ合う'), 0.5187432169914246)		(('軒下', '鏡板'), 0.5879324674606323)	
('鏡板', 0.4933130443096161)		(('縁側', '岐伯'), 0.5141000151634216)		(('囲炉裏', '注連縄'), 0.586726725101471)	
('板敷', 0.4912448823451996)		(('腰掛', '及時'), 0.5056247115135193)		(('明障子', '板敷き'), 0.5859760046005249)	
('葺き', 0.4908360242843628)		(('障子', '8074'), 0.5027346014976501)		(('桟', '注連縄'), 0.5852824449539185)	
('床の間', 0.4900680184364319)	正解	(('床の間', 'ハラルド・ゴルムスソン'), 0.4974367916584015	正解	(('連子窓', '板敷き'), 0.5843095779418945)	
('遣戸', 0.4891408681869507)		(('庇', 'iNaturalist'), 0.49132341146469116)		(('囲炉裏', '腰掛'), 0.5834441781044006)	
('腰掛', 0.4884079694747925)		(('鏡板', '飲める'), 0.4468616545200348)		(('連子窓', '明障子'), 0.5834190249443054)	
('しめ縄', 0.485820472240448)		(('葺き', 'シェアリング'), 0.3146193027496338)		(('囲炉裏', '遣戸'), 0.5830212235450745)	正解
('庇', 0.48550713062286377)		(('衝立', '='), 0.15735812485218048)		(('土間', '葺き'), 0.5826461911201477)	

「坐月子」

日本語類似単語	正解・不正解	類義語と差分ベクトルに一番近い日本語のペア	正解・不正解	日本語類義語のペア	正解・不正解
('授乳', 0.5040315389633179)	正解	(('授乳', 'ベスパニョーラ'), 0.5057529807090759)	正解	(('授乳', '発病'), 0.5665010809898376)	正解
('妊娠', 0.47521308064460754)		(('疳', '試着'), 0.4923381805419922)		(('妊娠', '入浴'), 0.564978837966919)	
('分娩', 0.45637047290802)		(('妊娠', 'ラバーストラップ'), 0.474890798330307)		(('授乳', '受診'), 0.5529268383979797)	
('入浴', 0.44373446702957153)	正解	(('分娩', 'アーバンソリューションズ'), 0.46194377541542053)		(('入浴', '発病'), 0.547091543674469)	
('発病', 0.4431608319282532)		(('入浴', 'アントニー・フォッカー'), 0.44417527318000793)	正解	(('入浴', '出産'), 0.5458847880363464)	正解
('婦人病', 0.4283590018749237)	正解	(('発病', 'ScreenX'), 0.4433394968509674)		(('授乳', '歯磨き'), 0.5451692938804626)	
('月経', 0.42809346318244934)		(('婦人病', 'アンツィフェーロフ'), 0.44173434376716614)	正解	(('授乳', '入浴'), 0.5446174740791321)	
('服用', 0.4259428083896637)		(('産後', '鋳込む'), 0.4409390687942505)	正解	(('授乳', '婦人病'), 0.5431270003318787)	正解
('歯磨き', 0.42572319507598877)		(('帝王切開', 'あけくれる'), 0.4405152499675751)		(('授乳', 'わずらっ'), 0.5423293113708496)	
('出産', 0.4255729615688324)	正解	(('つわり', 'WikiEducator'), 0.4374088644981384)		(('分娩', '入浴'), 0.540971040725708)	正解
('産後', 0.42445963621139526)	正解	(('用便', 'リオクリュー'), 0.4321373403072357)		(('婦人病', '用便'), 0.5398937463760376)	正解
('用便', 0.42272236943244934)		(('歯磨き', 'スダ・ドアカワールド'), 0.4296853244304657)		(('授乳', '産後'), 0.538608968257904)	正解
('診察', 0.42161691188812256)		(('月経', 'ソービー'), 0.42776742577552795)		(('入浴', '不妊'), 0.5383350849151611)	
('帝王切開', 0.420818388462066665)		(('出産', 'グループマテリアルズ'), 0.42527586221694946)	正解	(('妊娠', '歯磨き'), 0.5376395583152771)	
('受診', 0.4205842614173889)		(('診察', 'ロバート・トリヴァース'), 0.4236065745353699)		(('妊娠', '受診'), 0.5369066596031189)	
('疳', 0.41810041666030884)		(('受診', 'ダントレーヴ'), 0.42182308435440063)		(('妊娠', '診察'), 0.5367541909217834)	
('不妊', 0.41559314727783203)		(('流産', 'ベイラインエクスプレス'), 0.4144819974899292)	正解	(('授乳', '不妊'), 0.5361728668212891)	
('流産', 0.4148959219455719)	正解	(('不妊', '花文字'), 0.4139632284641266)		(('発病', '歯磨き'), 0.5350319147109985)	
('つわり', 0.4142608046531677)		(('わずらっ', '大皿'), 0.38194841146469116)		(('授乳', '用便'), 0.5342364311218262)	
('わずらっ', 0.4122973382472992)		(('服用', 'に際し'), 0.2556542754173279)		(('授乳', '妊娠'), 0.533598780632019)	正解

「交杯酒」

日本語類似単語	正解・不正解	類義語と差分ベクトルに一番近い日本語のペア	正解・不正解	日本語類義語のペア	正解・不正解
('手料理', 0.4311058223247528)		(('たらふく', '聴会'), 0.4551336169242859)	正解	(('手料理', '神酒'), 0.49598827958106995)	正解
('御飯', 0.42789560556411743)		(('御飯', 'ストリートボーラー'), 0.43829429149627686)		(('くだもの', '酒食'), 0.4955103397369385)	正解
('一杯', 0.42182597517967224)	正解	(('まろうど', '竜眼肉'), 0.4320698380470276)		(('手料理', '桃の節句'), 0.49477770924568176)	
('赤飯', 0.41902777552604675)		(('赤飯', 'Staffing'), 0.42545852065086365)	正解	(('手料理', '湯のみ'), 0.49267005920410156)	
('ご馳走', 0.4132000207901001)		(('バーバ・ヤーガ', '深から'), 0.4242188036441803)		(('一杯', '桃の節句'), 0.49238449335098267)	
('神酒', 0.40300726890563965)	正解	(('桃の節句', '誠'), 0.42209622263908386)		(('スンニュン', 'まろうど'), 0.4894832372665405)	
('柚子', 0.40002188086509705)		(('一杯', 'インターフェイスデザイン'), 0.41854527592658997)	正解	(('手料理', '朝露'), 0.48807939887046814)	
('スンニュン', 0.3956157863140106)		(('柚子', 'ユーゴスラビア・ドルガ・リーガ'), 0.4055337607860565)		(('手料理', '御神酒'), 0.4878701865673065)	正解
('湯のみ', 0.3878569006919861)		(('神酒', 'KOEVOET'), 0.40018829703330994)	正解	(('湯のみ', 'まろうど'), 0.486588716506958)	
('御神酒', 0.38697999715805054)	正解	(('朝露', '戒和'), 0.3931575119495392)		(('一杯', '酒食'), 0.4852534234523773)	正解
('くだもの', 0.38154685497283936)		(('スンニュン', 'アンジェリカ'), 0.38740402460098267)		(('手料理', 'くだもの'), 0.48350128531455994)	
('餅', 0.38141509890556335)		(('客人', '%')) 、'), 0.37821435928344727)		(('手料理', '一杯'), 0.48167848587036133)	
('桃の節句', 0.3814108073711395)		(('とろろ', 'ルクワイア'), 0.3764607310295105)		(('手料理', 'まろうど'), 0.4813903272151947)	
('たらふく', 0.3803858160972595)		(('餅', 'バーベリ'), 0.37154287099838257)		(('一杯', '神酒'), 0.4800145924091339)	正解
('まろうど', 0.3783974349498749)		(('湯のみ', '[...]'), 0.29130813479423523)		(('一杯', 'まろうど'), 0.4792499840259552)	
('バーバ・ヤーガ', 0.37544143199920654)		(('くだもの', '立ち会い'), 0.2399655580520063)		(('手料理', 'バーバ・ヤーガ'), 0.4787680506706238)	
('酒食', 0.3748187720775604)	正解	(('ご馳走', 'ならびに'), 0.15680374205112457)		(('手料理', '御飯'), 0.4779466688632965)	
('客人', 0.37462466955184937)		(('手料理', 'BR'), 0.10776444524526596)		(('柚子', '客人'), 0.4775286614894867)	
('朝露', 0.37120136618614197)		(('御神酒', '率いる'), 0.0941191241145134)	正解	(('まろうど', '朝露'), 0.4763644337654114)	
('とろろ', 0.3708658814430237)		(('酒食', '-'), 0.06196361407637596)	正解	(('手料理', '柚子'), 0.4757183790206909)	

「数九」

日本語類似単語	正解・不正解	類義語と差分ベクトルに一番近い日本語のペア	正解・不正解	日本語類義語のペア	正解・不正解
('句', 0.4314446449279785)		(('句', 'NBTC'), 0.4320239722728729)		(('公案', '太陰暦'), 0.5303564071655273)	
('公案', 0.4231099486351013)		(('公案', 'ポールウインナー'), 0.4232088029384613)		(('公案', '陰暦'), 0.5295858979225159)	
('劫', 0.41505634784698486)		(('六経', 'サリー・コンウェイ'), 0.4173288643360138)		(('公案', '節気'), 0.5238650441169739)	
('干支', 0.41472306847572327)	正解	(('五行', 'ウィーゲル'), 0.4130459129810333)		(('公案', '立春'), 0.5237380266189575)	
('賦', 0.41414421796798706)	正解	(('干支', 'エブスコ'), 0.41117122769355774)	正解	(('劫', '干支'), 0.5237055420875549)	
('五行', 0.4102438688278198)		(('賦', 'デ・ルーカ'), 0.408670037984848)	正解	(('賦', '冬至'), 0.5205201506614685)	正解
('六経', 0.40801912546157837)		(('冬至', 'エブスコ'), 0.4035686254501343)	正解	(('干支', '賦'), 0.518984317779541)	正解
('冬至', 0.4072076380252838)	正解	(('太陰暦', '14200'), 0.40088847279548645)	正解	(('劫', '太陰暦'), 0.5186171531677246)	
('発句', 0.40264809131622314)		(('〃', 'コ・フェスタ'), 0.3996896743774414)		(('冬至', '帖'), 0.518237292766571)	正解
('頌', 0.4022463858127594)	正解	(('陰暦', 'チュンソフトプロジェクト'), 0.3987753093242645)	正解	(('句', '冬至'), 0.5147366523742676)	正解
('帖', 0.3933136761188507)		(('節気', '6061'), 0.384641170501709)	正解	(('公案', '干支'), 0.5125520825386047)	
('陰暦', 0.39154586195945574)	正解	(('立春', '6061'), 0.3841356039047241)	正解	(('公案', '冬至'), 0.5121886134147644)	
('太陰暦', 0.39115706086158750)	正解	(('発句', 'ケビン・ラッド'), 0.38103777170181274)		(('劫', '立春'), 0.5102246999740601)	
('律詩', 0.3845991790294647)	正解	(('二十八宿', 'ベルナッキ'), 0.38098493218421936)		(('句', '太陰暦'), 0.5095160007476807)	正解
('節気', 0.3797317147254944)	正解	(('韻', 'ヘルス・エンジェルス'), 0.3745937943458557)		(('句', '干支'), 0.5093138217926025)	
('立春', 0.37845155596733093)	正解	(('律詩', 'リバティメディア'), 0.340493381023407)	正解	(('干支', '頌'), 0.5072511434555054)	
('二十八宿', 0.3781031370162964)		(('頌', 'CTU'), 0.3033263683319092)		(('句', '劫'), 0.5065075159072876)	
('卦', 0.3774862289428711)		(('帖', 'クレイ'), 0.2507103979587555)		(('頌', '太陰暦'), 0.5064473152160645)	
('〃', 0.37645280361175537)		(('卦', '&'), 0.19596172869205475)		(('冬至', '頌'), 0.5063366293907166)	正解
('韻', 0.37633803486824036)		(('劫', '地元'), 0.11726463586091995)		(('劫', '節気'), 0.5058206915855408)	

「娃娃亲」

日本語類似単語	正解・不正解	類義語と差分ベクトルに一番近い日本語のペア	正解・不正解	日本語類義語のペア	正解・不正解
('養母', 0.5201137065887451)		(('養母', 'Peten'), 0.5216278433799744)		(('養母', '初婚'), 0.5577341914176941)	
('連れ子', 0.4840914011001587)		(('連れ子', '魚形水雷'), 0.4841614067554474)		(('養母', '間の子'), 0.545772135257721)	
('義母', 0.4739173948764801)		(('義母', 'ポペットバルブエンジン'), 0.47812914848327637)		(('養母', '連れ子'), 0.5446956157684326)	
('義妹', 0.4731845259666443)		(('義妹', 'ワケルンジャー'), 0.4724116623401642)		(('養母', '前夫'), 0.5384843945503235)	
('先妻', 0.46938610076904297)		(('先妻', 'ワケルンジャー'), 0.46903958916664124)		(('養母', '許嫁'), 0.5355861783027649)	
('実母', 0.468101711008110046)		(('実母', 'ルジャーナ'), 0.4642825722694397)		(('養母', '腹違い'), 0.5353318452835083)	
('許嫁', 0.4629703164100647)	正解	(('後妻', 'TPC'), 0.4618526101112366)		(('養母', '前妻'), 0.5342666506767273)	
('後妻', 0.4619660973548889)		(('義姉', 'ネイティブペタ'), 0.45347851514816284)		(('養母', '先妻'), 0.5342534184455872)	
('兄嫁', 0.46164053678512573)		(('舅', 'CLEAN'), 0.4503838121891022)		(('養母', '義妹'), 0.5336127281188965)	
('前妻', 0.4577442705631256)		(('妹', 'CTF'), 0.44779306650161743)		(('養母', '兄嫁'), 0.5334611535072327)	
('間の子', 0.4548311233520508)		(('伯母', 'シノビカード'), 0.44118064641952515)		(('養母', '愛娘'), 0.5234794020652771)	
('前夫', 0.453483521938324)		(('腹違い', 'ワケルンジャー'), 0.4342552721500397)		(('義妹', '間の子'), 0.5234087109565735)	
('義姉', 0.44893762469291687)		(('継母', '〞:『'), 0.42900246381759644)		(('養母', '舅'), 0.5232824087142944)	
('初婚', 0.4489002227783203)	正解	(('前妻', '教範'), 0.4065123498439789)		(('実母', '間の子'), 0.5229233503341675)	
('妹', 0.4460832476615906)		(('愛娘', 'CSG'), 0.34925076365470886)		(('養母', '義姉'), 0.5229011178016663)	
('継母', 0.4442523717880249)		(('間の子', 'まるごと'), 0.2283579260110855)		(('養母', '義母'), 0.5222693085670471)	
('舅', 0.44136443734169006)		(('初婚', 'あわせ'), 0.20073789358139038)	正解	(('許嫁', '初婚'), 0.5209137201309204)	正解
('愛娘', 0.44068577885627747)		(('前夫', 'まるごと'), 0.18081946671009064)		(('前妻', '舅'), 0.5203913450241089)	
('伯母', 0.4381445050239563)		(('兄嫁', '於け'), 0.08278096467256546)		(('初婚', '舅'), 0.5191904306411743)	
('腹違い', 0.4379097819328308)		(('許嫁', '及び'), 0.0800158903002739)	正解	(('義妹', '初婚'), 0.5179152488708496)	正解

付録：ソースコード

A.1 日本語類義語ごとペアの作成コード

```
from gensim.models import word2vec
import gensim
import math
from scipy import spatial
import numpy as np
import pandas as pd

#バイナリファイルなら True,txt ファイルだったら False
#中国語のモデル
model1 = gensim.models.KeyedVectors.load_word2vec_format('vectors-zh.bin', binary=True)
#日本語のモデル
model2 = gensim.models.KeyedVectors.load_word2vec_format('vectors-jp.bin', binary=True,unicode_errors='ignore')
#これで文化固有単語のベクトルを取り出す

cul_data=model1["汤圆"]


#これで日本語の類義語 20 個を取り出す
lui_jp = model2.most_similar(positive=[cul_data], topn=20)
print("日本語類似単語")

#類義語の組み合わせ
import itertools
#20 個日本語類義語を全組み合わせで 190 のペアを作る
pa_li=list(itertools.combinations(lui_jp_li,2))
# print(pa_li)
ra_li=[]
```

```
for word in pa_li:

    lui_1=word[0]
    #print(lui_1)
    lui_2=word[1]
    #print(lui_2)
    #各ペアで単語ベクトルを足し算してできた単語ベクトル
    plus_vec=model2[lui_1]+model2[lui_2]
    #ランク値＝和ベクトルと文化固有単語ベクトルの類似度が計算
    ans=1-spatial.distance.cosine(cul_data,plus_vec)
    ra_li.append(ans)

p_r_li=[]
for p,r in zip(pa_li,ra_li):
    p_r=(p,r)
    p_r_li.append(p_r)
#p_r_li=ランク値付けの類義語のペアリスト
# print(p_r_li)

re_li=dict(p_r_li)
re_list = sorted(re_li.items(), key=lambda x:x[1],reverse=True)
#print(re_list)
#re_list=類義語のペア付けでランク値を降順でソートしたもの
```

A.2 類義語と差分のペアの作成コード

```
lui_jp_li=[]
clo_jp_li=[]
rank_li=[]
c_j_l=[]
#類義語と差分の組み合わせ

for word in lui_jp:
```

```
    #日本語類義語 20 個を取り出す
    lui_jp_li.append(word[0])
    #これで差分ベクトル＝文化固有単語ベクトルー日本語類義語単語ベクトルを取り出す
    sabun_vec=cul_data-model2[word[0]]
    #print(sabun_vec)
    #差分ベクトルに一番近い日本語の単語ベクトルを取り出す
    clo_jp=model2.most_similar(positive=[sabun_vec],topn=1)

    for Word in clo_jp:
        #一番近い日本語の単語ベクトル+日本語類義語単語ベクトル＝文化概念ベクトル
        lui_zh=model2[Word[0]]+model2[word[0]]
        #文化概念ベクトルと文化固有単語ベクトルの類似度が計算
        cos_ans=1-spatial.distance.cosine(cul_data,lui_zh)
        rank_li.append(cos_ans)#一番近い日本語単語と日本語類似単語のランク値
        #文化固有単語–日本語類似単語に一番近い日本語単語
        clo_jp_li.append(Word[0])
        c_j_l.append(Word)
#print(c_j_l)
pair_li=[]
pa_ra_li=[]
for w1,w2 in zip(lui_jp_li,clo_jp_li):
    pair=(w1,w2)
    pair_li.append(pair)
for pair1,rank1 in zip(pair_li,rank_li):
    pair_rank=(pair1,rank1)
    pa_ra_li.append(pair_rank)

#pa_ra_li=日本語類義語と一番近い日本語単語をペアでランク値付けしたもの
```

```
result_li=dict(pa_ra_li)
result_list = sorted(result_li.items(), key=lambda x:x[1],reverse=True)
#result_list=類義語と差分で取得したペア付けでランク値を降順でソートしたもの
```