情報処理学会論文誌 Vol.55 No.1 257–266 (Jan. 2014)

日本語版・中国語版Wikipediaを用いた文化差検出手法の提案

諏訪 智大[1,a)] 宮部 真衣[2,b)] 吉野 孝[1,c)]

受付日 2013年4月10日, 採録日 2013年10月9日

概要：多言語間コミュニケーションにおいて，同一の単語を用いて会話をしている場合でも，相手の文化について十分に理解していないために，誤解が生じる可能性がある．現在，文化差の有無の判断は，人が行う必要があるが，その判断には相手の文化に関する十分な知識が必要となるため，容易ではない．そのため，文化差が存在することを自動的に検出する仕組みが求められている．そこで本論文では，多言語知識のデータベースであるWikipediaを利用した文化差の検出手法を提案する．検討用データセットを分析した結果，文化差のある語句には，Wikipedia記事の本文やカテゴリに特徴が存在することが分かった．そこで，これらの特徴を考慮した手法を検討した．実験の結果，提案手法は，検討用データセットにおけるすべての文化差に関して，最も良い精度で検出することができた．また，「世界的な観点から説明されていない」「日本語版Wikipediaのみで日本について言及されている」「中国語版Wikipediaにおいて中国について偏った表記がされていない」「各言語版Wikipedia内に各国に対する言及表現が存在しない」の特徴を含む記事が，文化差判定において重要であることが分かった．

キーワード：文化差，多言語間コミュニケーション，Wikipedia

Proposal of Cultural Difference Detection Method Using Japanese and Chinese Wikipedia

Tomohiro Suwa[1,a)] Mai Miyabe[2,b)] Takashi Yoshino[1,c)]

Received: April 10, 2013, Accepted: October 9, 2013

Abstract: There is a possibility that the misunderstanding is caused in multilingual communications, because people cannot understand enough other culture even when talking by using the same word. People should judge the presence of a cultural difference in today's situation. However, it is difficult to judge it because enough knowledge of other culture is necessary. We proposed a cultural difference detection method using Wikipedia which is a database of multilingual knowledge. From the analysis of data set for examination, We found the feature in the text and category, and others of the articles on Wikipedia. Therefore, We proposed method using these features. From the result of the experiment using data set for examination, we presented that the proposed method can detect correct differences about all kind of cultural differences. We found that the following features of an article are important for detecting cultural difference: (1) An article is not explained from a global viewpoint. (2) An article is mentioned about Japan by Japanese version Wikipedia. (3) An article of the Chinese version Wikipedia is not mentioned which inclined toward China. (4) An article in each language version Wikipedia has not mentioned each country.

Keywords: cultural difference, multilingual communications, Wikipedia

1. はじめに

多言語間コミュニケーションにおいて，同一の単語を用いて会話をしている場合でも，相手の文化について十分に理解していないために，誤解が生じる可能性がある [1]．これまでに，遠隔チャット中や対面コミュニケーション中に，

[1] 和歌山大学
Wakayama Uniersity, Wakayama 640–8510, Japan
[2] 京都大学
Kyoto University, Kyoto 600–8815, Japan
[a)] s145019@sys.wakayama-u.ac.jp
[b)] mai.miyabe@gmail.com
[c)] yoshino@sys.wakayama-u.ac.jp

画像などのアノテーションを付与する手法を用いて，誤解を減らす工夫がなされている [2], [3]．しかし，アノテーションの付与が必要となる語句の選択は，利用者自身が行う必要があった．つまり，文化差の有無の判断は，利用者自身が行う必要がある．しかし，その判断には相手の文化に関する十分な知識が必要となるため，容易ではない．そのため，文化差が存在することを自動的に検出する仕組みが求められている．

本論文では，多言語知識のデータベースである Wikipedia を利用した文化差の検出手法を提案する．まず検討用データセットとして文化差があると感じられる語句 114 語句を収集し，そこから文化差を判定する特徴を分析し，文化差検出手法を提案する．実験として，検討用データセットに対する文化差検出精度を評価し，それぞれの手法を分析する．さらに，提案する手法が，Wikipedia からランダムに選択された語句に対しても有効かどうかを検証する．

2. 関連研究

関連研究として，まず，異文化間コミュニケーションにおける，文化差に関する研究を示す．Cho らは，異文化話者らがコンピュータとネットワークを介してコミュニケーションを行う際に用いる絵文字に着目した．絵文字は，異文化間で普遍的に解釈されないという問題がある．そこで，その問題を解決するために，解釈に文化差のある絵文字の検出における工学的な手法の適用可能性について検討した [4]．検討の結果，従来の工学的な手法では，人の文化差判定を近似することは困難であることを示した．Koda らは，アバタを介したコミュニケーションにおける，異文化間での表情の解釈に着目した．アバタの表情に関するユーザの解釈について実験を行った結果，表情の解釈が文化によって大きく異なることを示した [5]．文化差に関しては，これまでにいくつかの検討が行われているが，文化差判定は容易ではない．

Wikipedia の複数の言語版を利用した研究を示す．Pfeil らは，Hofstede が明らかにした，国ごとに存在する 4 つの文化的多様性 [6] を Wikipedia に適用し，それぞれの国ごとの Wikipedia の編集操作は，それぞれの国の文化的特徴と相関することを明らかにした [7]．藤原らは，自国の言語版の Wikipedia だけでは情報量が不足する場合の補完のために，多言語版の Wikipedia に対して，リンク構造解析を用いることで，差異情報を抽出する方法を提案している [8]．松浦らは，日本語と外国語での同一ニュースに関する変遷を分析するために，Wikipedia を用いている [9]．吉岡は，機械翻訳システムの精度向上のために，Wikipedia の言語間リンクを用いた中日の翻訳辞書の作成方法を提案している [10]．

このように，Wikipedia は知識抽出分野で資源として注目を集めており，様々な利用が検討されている．しかし，これまでに，Wikipedia の多言語データを利用した文化差検出に関する試みは行われていない．

3. 文化差の定義

本章では，本研究で検出対象とする「文化差」について定義する．

文化差を定義するためには，まず，「文化」の定義が必要である．「文化」（Culture）の定義は，日本と欧米では異なり，一概に定義することは困難である [11]．たとえば，今日欧米で用いられる「文化」は，「知識，信仰，芸術，道徳，慣習，その他社会の一員としての人間によって獲得される能力や習慣を包含する複合体である」と定義づけられている [11]．このような「文化」を単純に「測る」ことは困難であるが，コミュニケーション支援に文化差検出手法を適用するためには，何らかの尺度を考える必要がある．そこで本論文では，特に「知識」の面から「文化」をとらえることとし，まず，文化差検出手法の第 1 歩として，形式知化された知識の違いで文化差を測ることとした．

次に，「第 1 種の文化差」と「第 2 種の文化差」を定義する．「第 1 種の文化差」のある内容は，一方の文化圏で発生したり，存在したりしている「もの」や「こと」で，基本的には，もとの文化圏の内容を指しているが，伝わっている知識が限定的であり，もとの文化圏における解釈が完全には再現されないものである．たとえば，日本の地名のいくつかは，海外にも伝わっているが，その地名の持つ背景（歴史的あるいは文化的）などは正確には理解されない．「第 2 種の文化差」のある内容は，どちらの文化圏にも存在するが，それぞれの文化圏で意味の異なるものである．たとえば，「醤油」は日本と中国のどちらにも存在するが，日本の「醤油」と中国の「醤油」は異なる．

4. 文化差の検出手法

4.1 検討用データセットの作成

文化差検出手法の検討にあたり，まず，文化差のある語句の分析を行う．そこで，分析対象とする検討用データセットを作成した．検討用のデータセットは次の手順で作成した．

Step 1 本学の学生 18 名に依頼し，下記に該当する日本語の語句を収集した．

- 日本独特のもの（と思っている語句）
- 日本にも海外にもあるが，日本と海外とは違う（と思っている語句）

ただし，地名，人名，固有名詞は，ほとんど第 1 種の文化差のある語句に該当するため，入力しないように依頼した．

Step 2 収集した語句が Wikipedia の日本語版と中国語版の両方に，記事として存在するか調べた．

Step 3 日本に 3 年以上住んでいる本学の中国人留学生 5

表 1 データセットの一部と留学生による分類結果
Table 1 A part of experiment data set and decision results by foreign students.

ID	語句	分類結果（2 名以上）	第 1 種の文化差	第 2 種の文化差	日本と中国の違いについて
4	外国人	文化差なし			—
6	干支	日本とは違う		○	日本では，亥（いのしし）だが，中国では豚．
28	刺身	中国にはない	○		中国では，あまり生の食べ物は食べない．
45	煎餅	日本とは違う		○	中国では煎餅は主食，日本ではお菓子．
49	ケーキ	文化差なし			—
53	天津飯	中国にはない	○		中華料理に，ご飯の上に卵をのせる料理はない．
57	ラーメン	日本とは違う		○	日本のラーメンとは味が違う．
79	観光	文化差なし			—
85	ゲーム	文化差なし			—
100	パチンコ	中国にはない	○		中国ではギャンブルは禁止されている．
107	映画	文化差なし			—
108	納豆	中国にはない	○		中国に納豆はない．
112	醤油	日本とは違う		○	中国の醤油は塩辛い．
113	おにぎり	中国にはない	○		中国では，冷めたご飯は食べない．
114	饅頭	日本とは違う		○	日本の饅頭の中には，餡がある．

名に依頼し，Step 1 で収集した語句のうち Wikipedia に記事（日本語版と中国語版の両方）として存在するもの（Step 2 で調べた結果によって抽出したもの）を，次の 3 種類に分類してもらった．

- 中国にはない（第 1 種の文化差）．
- 中国にもあり，日本と同じ．
- 中国と日本のものは違う（第 1 種の文化差または第 2 種の文化差）．

「中国と日本のものは違う」を選択した場合には，どのように違うかを簡単に記述してもらった．

Step 1 で，日本の語句が，2,200 語句以上集まった．Step 2 で，入力語句の 5%程度について，入力された順に Wikipedia の記事の有無を調べ，114 語句をデータセットとした．

表 1 に，今回利用したデータセットの一部と留学生による分類結果を示す．語句の分類は，分類作業を行った各中国人留学生の経験に基づいた意見をもとにしているため，事実と異なる可能性がある．また，表に示した分類結果は，5 名の留学生による分類結果を集計し，最も多い分類結果（2 名以上が分類）を代表値とした場合の結果である．「日本と中国との違いについて」の内容は，留学生らの記述をもとにしている．第 1 種の文化差および第 2 種の文化差の分類は，著者らの定義に基づき，留学生らの記述や Wikipedia の記述を参考に，著者らが分類した．なお定義に基づく著者らの分類結果は，全員で一致したため，その結果を採用している．

4.2 文化差判定のための指標

4.1 節で作成した検討用データセットを用いて，文化差のある語句・ない語句における，文化差判定に有用である可能性がある特徴について分析した．確認の結果，記事の Wikipedia 独自の構造的な特徴と，記事内に共通して現れる特徴が見られた．本節では，分析によって明らかになった，文化差判定において有効な指標となりうる Wikipedia 記事における特徴をまとめる．指標としては，特定の語句に関する特徴と，すべての語句に関する特徴の 2 つに分けることができる．以下に，特徴を示す．

- 特定の語句に関する特徴
  - 本文中の特徴
    - ＊ 世界的な観点から説明されていない記事
      「世界的な観点から説明されていないおそれがあります」という記述がある記事が「自動販売機」や「公衆便所」など 12 件存在する．
    - ＊ 日本の文化に関連した書きかけ項目である記事
      「浴衣」に「日本の文化に関連した書きかけ項目です」という記述が存在する．
  - カテゴリ中の特徴
    分析語句と同名のカテゴリが存在しており，そのカテゴリを説明する記事におけるカテゴリに「日本の」という記述が存在する語句が「忍者」や「禅」など 11 件存在する．
  - タイトルの特徴
    - ＊ 日本語版 Wikipedia において，「日本の」と記述されているタイトルの記事
      「日本のアニメ」や「日本のタクシー」など 17 件存在する．
    - ＊ 中国語版 Wikipedia において，「(日本)」と記述されているタイトルの記事
      「鬼（日本)」「祭（日本)」「団子（日本)」の 3 件存在する．
- すべての語句に関する特徴
  - 項目数の特徴
    一方の文化圏で発生したり，存在したりしているも

のについては，発生した文化圏とそれ以外の文化圏での知識量に違いがある可能性がある．このような場合，Wikipedia の記事内の項目数に反映される．

– 国名数・言語名数の特徴
ある語句の意味するものが文化によって異なる場合，各言語版の Wikipedia の記事の内容は，それぞれの文化に基づく内容になる可能性があると考えられる．また，他の文化圏から伝わった場合，そのような記述が含まれる可能性がある．そういった場合，記事内に国名や言語名が出現すると考えられる．

– 執筆者の意図の特徴
Wikipedia の記事においては，国や文化によって違いがある場合には，執筆者がそれぞれの文化の内容に関して，記述したり，カテゴリ分けしたりする場合がある．

これらの特徴は，文化差判定時に有効な指標となる可能性があると考えられる．

4.3 検出の仕組み

本章では，4.2 節で示した文化差判定のための指標を考慮した文化差の検出方法について説明する．**図 1** に提案する手法の流れを示す．

まず，提案手法に用いる判断の指標は，言及率などの記述内容を用いた指標と，「世界的な観点から説明されていないおそれがあります」といった記述内容に関する注釈[*1]などの語句の説明以外の指標，そしてタイトルの特徴などの Wikipedia の構造的な指標に分けることができる．提案手法を考慮するにあたり，より適切に文化差を判断することができると考えられる指標を用いた判断から適用していくことが好ましいと考えられる．Wikipedia の記述内容は，記述不足であったり，独自の研究が記述されていたりするため，文化差を判断する指標としては適切でない場合がある．よって，記述内容よりも，記述内容に関する注釈を優先する．また，記述内容に関する注釈よりも，Wikipedia の構造的な特徴のほうが，記述内容による執筆者の意図が入らないため有効であると考えられる．そこで，

(1) 構造的な特徴（判断 A，B）
(2) 記述内容に関する注釈（判断 C～E）
(3) 記述内容を用いた指標（判断 F～J）

の順に文化差の判断を適用していくことを考慮した．

判断 A：カテゴリにおける言及表現と同名カテゴリページにカテゴリ「日本の」の有無
まず，判断 A において，カテゴリ「日本の」の有無を調べる．ある記事の属する「カテゴリ」に「日本の」が含まれる場合には，何らかの文化的な内容が含まれていると見なした．また，記事名と同名であるカテゴリが存在し，そのカテゴリのページにカテゴリ「日本の」が存在する場合のどちらかを満たしていればよいものとする．条件を満たした場合，それ以降は第 1 種，または第 2 種の文化差のどちらの文化差であるかを考慮しつつ，判断 B から判断 H まで順に文化差を判断していくこととする．満たさなかった場合，文化差がな

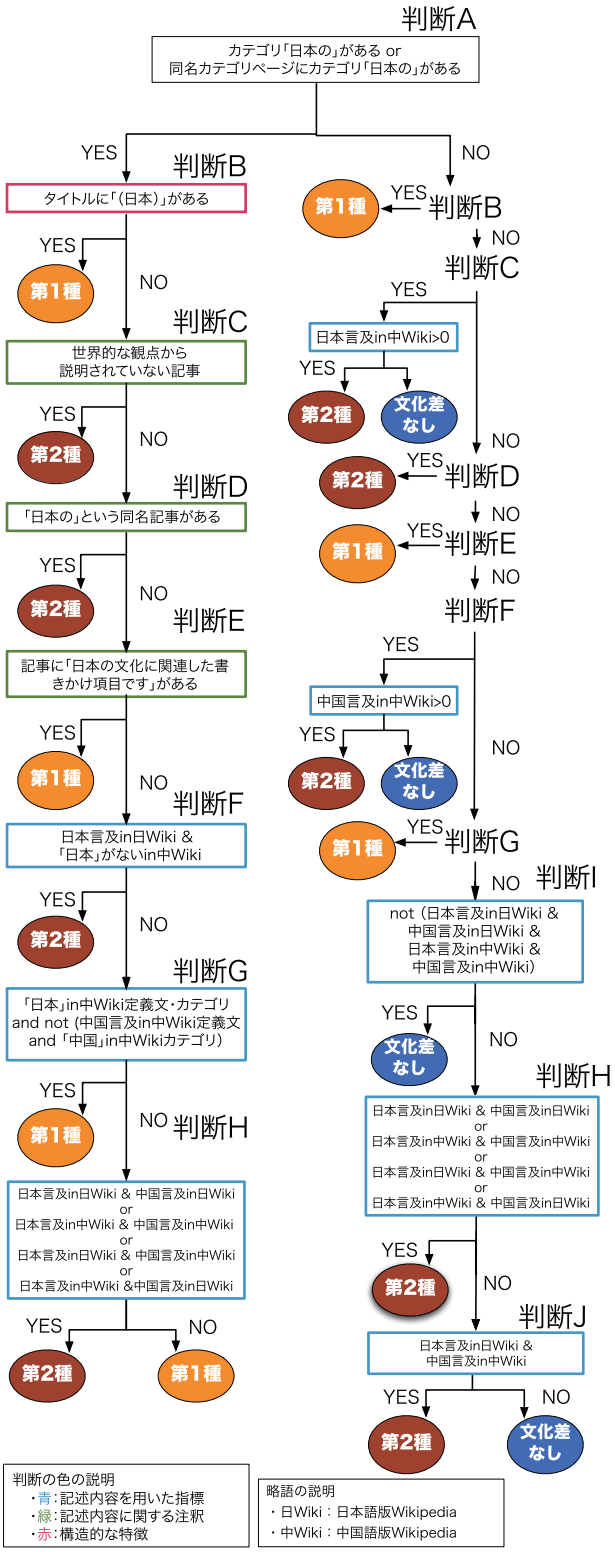


図 1 文化差検出の流れ
Fig. 1 Flowchart of cultural difference detection.

[*1] http://ja.wikipedia.org/wiki/Wikipedia:Template_メッセージ一覧/問題のある記事

いことを考慮しつつ，判断 B から判断 G，そして判断 I，H，J の順に文化差の判断を進めていくこととする．

判断 B：タイトルにおける国名表記の有無

まず，Wikipedia の構造的な特徴として，中国語版 Wikipedia の記事のタイトルにおいて，「(日本)」という表記が存在する場合，日本特有のことに関して説明している記事であり第 1 種の文化差と判定する．そして，以降は，Wikipedia の記述内容について書かれた判断を適用する．

判断 C：世界的な観点から説明されていない記事

日本語版 Wikipedia において，「世界的観点からの説明がされていないおそれがあります」という表記が存在する場合，その記事は本来一般的な項目であるが，日本について偏った説明をしており，第 1 種の文化差ではない可能性がある．よって，判断 A の条件を満たした場合，第 2 種の文化差と判定されることになる．また判断 A の条件を満たしておらず，中国語版 Wikipedia で日本について言及している場合，日本発祥ではないが何らかの文化差が存在すると考えられるため，第 2 種の文化差と判定し，そうでなければ文化差なしと判定する．

判断 D：「日本の」という記事の有無

日本語版 Wikipedia において，「詳細は日本の～を参照」などの記述が存在したり，タイトルに「日本の」という表記で別に記事が存在している場合，本来一般的な項目であるが，日本独自の説明をする記事も存在しており，第 2 種の文化差と判定する．

判断 E：日本の文化に関連した書きかけ項目である記事

日本語版 Wikipedia において，「日本の文化に関連した書きかけ項目です」という表記が存在する場合，その記事は日本発祥である語句を説明していると考え，第 1 種の文化差と判定する．

判断 F：日本語版 Wikipedia で日本に言及しているが，中国語版 Wikipedia で「日本」という表記が存在しない場合

日本語版 Wikipedia で自国について言及していても，中国語版 Wikipedia で「日本」という語句が出現しない場合，第 1 種の文化差ではないと考え，第 2 種の文化差と判定する．判断 A の条件を満たしていない場合は，中国語版 Wikipedia で中国について言及している場合，自国について説明しており，何らかの文化差があると考えられるため第 2 種の文化差と判定し，そうでなければ文化差なしと判定する．

判断 G：定義文とカテゴリにおける国名数と言及表現の有無

中国語版 Wikipedia の定義文またはカテゴリにおいて，「日本」という表記が存在し，定義文における中国への言及表現と，カテゴリにおける「中国」という表記が存在しないとき，中国語版 Wikipedia であるのに自国に偏った表記がされておらず，日本中心に記事が書かれており，第 1 種の文化差と判定する．

判断 H：各言語版 Wikipedia 内の言及率

判断 H では，カテゴリにおける言及表現が存在した語句に対しては，日本語版および中国語版 Wikipedia における日本言及率および中国言及率に基づき，第 1 種または第 2 種の文化差と判定する．ここで，日本言及率は，言及表現が存在する場合，0 以上となる．判断 H においては，以下の 4 つの条件のうち，いずれかを満たす場合に「第 2 種の文化差」と判断する．

(1) 日本言及 in 日 Wiki>0 & 中国言及 in 日 Wiki>0
(2) 日本言及 in 中 Wiki>0 & 中国言及 in 中 Wiki>0
(3) 日本言及 in 日 Wiki>0 & 中国言及 in 中 Wiki>0
(4) 日本言及 in 中 Wiki>0 & 中国言及 in 日 Wiki>0

上記の条件が成り立たない場合には，自国あるいは相手国の記述のみであるか，言及が存在しないため，「第 1 種の文化差」と判断する．また，判断 A の条件を満たしていない場合は，次の判断へ進む．

判断 I：各言語版 Wikipedia 内に言及表現がない

各言語版 Wikipedia 内に，日本および中国に関する言及がない場合は，「文化差なし」と判断する．

判断 J：各国の Wikipedia における自国への言及の有無

日本語版 Wikipedia において日本について言及があり，中国語版 Wikipedia において，中国に言及している場合を，ともに満たす場合は第 2 種の文化差と判定し，そうでなければ文化差なしと判定する．

5. 実験

5.1 実験条件

4.2 節で示した指標のうち，それぞれ 1 つの指標のみで，文化差を判断することができる手法と提案手法を比較検討しつつ，提案手法の文化差検出精度を適合率[*2]，再現率[*3]および F 値[*4]を用いて検証する[*5]．比較する手法を以下に示す．

比較手法 1 項目数に基づく手法

一方の文化圏で発生したり，存在したりしているものについては，発生した文化圏とそれ以外の文化圏での知識量に違いがある可能性がある．このような場合，Wikipedia の記事内の項目数に反映されるのではないかと考えられる．そこで，項目数の差に基づき第 1 種の文化差を検出する．項目数の差をもとに文化差を判定する場合，文化差があると判断する基準が必要とな

*2 判定結果が，正解データとどの程度一致しているかを表す．
*3 判定結果が，正解データをどのくらい網羅しているかを表す．
*4 「適合率」と「再現率」の調和平均を表す．
*5 本論文では，「第 1 種の文化差」「第 2 種の文化差」「文化差あり（第 1 種 + 第 2 種の文化差）」「文化差なし」をそれぞれ検出対象とした場合の精度を検証することとし，それぞれの適合率，再現率，F 値を算出する．

る．そこで，文化差があると考えられる語句として，「日本の観光地名（以下，観光地名）」[*6]（211 語句）を，文化差がない（少ない）と考えられる語句として，「すべての言語版にあるべき項目の一覧（以下，優先項目）」[*7]（1000 語句），「コンピュータ用語一覧（以下，コンピュータ用語）」[*8]（457 語句）を用いて基準を検討した．文化差があると考えられる「観光地名」については，項目数の差は平均 74.8%（標準偏差 31.8%）という結果になり，さらに項目数の差が 50%以上のものは，地名全体の 81%程度となった．一方，文化差がない（少ない）と考えられる「優先項目」「コンピュータ用語」については，それぞれ平均 49.7%（標準偏差 32.0%）の違いが見られた．つまり，文化差がないと考えられる語句についても，50%程度の項目数の違いは発生しうると考えられる．そこで，比較手法 1 では暫定的に 50%を判定基準として用いることとした[*9]．

比較手法 1 による各文化差検出の概要を以下に示す．

- **第 1 種の文化差の検出：**Wikipedia における記事の項目数を利用する．項目数に 50%以上の違いがある場合，「第 1 種の文化差」があると判定する．
- **第 2 種の文化差の検出：**比較手法 1 は，第 1 種の文化差の検出を目的とした手法であるため，第 2 種の文化差の検出はできない．

比較手法 2 記事内の国名・言語名数に基づく手法

ある語句の意味するものが文化によって異なる場合，各言語版での Wikipedia の記事の内容は，それぞれの文化に基づく内容になる可能性があると考えられる．また，他の文化圏から伝わった場合，そのような記述が含まれる可能性がある．そういった場合，記事内に国名や言語名が出現すると考えられる．そこで，国名・言語名数の差に基づき，第 1 種および第 2 種の文化差を検出する．国名・言語名数の差についても，比較手法 1 の基準値検討に用いたデータにより，文化差があると判定する基準について検討したデータを確認した結果，文化差のある語句については，特定の言語名が多くなる傾向が見られたため，暫定的に設定した[*10]．

比較手法 2 による各文化差検出の概要を以下に示す．

- **第 1 種の文化差の検出：**異なる言語版の Wikipedia の記事において，ある特定の国名・言語名が 1 件でも多い場合には，各記事は，ある特定の国名・言語名に関する説明を行っている記事であり，その国名・言語名の文化に属する内容であり「第 1 種の文化差」があると判定する．
- **第 2 種の文化差の検出：**記事に含まれる国名・言語名の数を利用する．どちらの記事にも，ある特定の国名・言語名が 1 件でも多い場合には，各記事は，同じ内容の説明を行っている記事であり，「文化差がない」と判定する．逆に，各言語版の国名・言語名が 1 件でも多い場合は，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 1 件でも多い場合には，「第 2 種の文化差」があると判定する．

比較手法 3 記事における執筆者の意図に基づく手法

Wikipedia の記事においては，国や文化によって違いがある場合には，執筆者がそれぞれ文化の内容に関して，記述したり，カテゴリ分けしたりする．そこで，このような執筆者の意図に基づき，第 1 種および第 2 種の文化差を検出する．なお，本手法では，日本語版の Wikipedia のみを利用している．比較手法 3 による各文化差検出の概要を以下に示す．

- **第 1 種の文化差の検出：**Wikipedia における各記事のカテゴリとして，日本に関するカテゴリを選択している場合に，「第 1 種の文化差」があると判定する．具体的なカテゴリとしては，「日本の食文化」「日本の年中行事」などのカテゴリがある．
- **第 2 種の文化差の検出：**各記事内の記述として，「日本では」「中国では」という記述をそれぞれ検索する．それぞれの検索数が，1 件以上の場合に，その記事は，日本と中国に関して記述を行っている記事であり，「第 2 種の文化差」があると判定する．

なお，第 1 種の文化差と第 2 種の文化差は，両方同時に検出される場合がある．第 2 種の文化差を検出している時点で，両方に存在しているものであると判断できるため，第 1 種の文化差と第 2 種の文化差のどちらも検出した場合は，第 2 種の文化差と判定する．

5.2 検討用データセットを用いた検出実験

4.1 節で示した 114 語句を用いて，5.1 節で示した各手法の文化差検出性能を比較する．114 語句中，「第 1 種の文化差」がある語句は 42 件，「第 2 種の文化差」がある語句は 41 件，「文化差なし」である語句は 31 件である．「第 1 種の文化差」「第 2 種の文化差」「文化差なし」という 3 種類に分類した場合の判定精度と，「文化差あり」「文化差な

[*6] http://www.jnto.go.jp/jpn/tourism_data/data_service.html
[*7] http://ja.wikipedia.org/wiki/Wikipedia:すべての言語版にあるべき項目の一覧
[*8] http://ja.wikipedia.org/wiki/コンピュータ用語一覧
[*9] 5.2 節で後述する検討用データセットに対して閾値を変更して精度を検証したところ，閾値を 60%としたときが最も精度が良く，次いで今回採用した 50%の精度が良いという結果が得られた．また，閾値 60%，50%の場合の精度に大きな差はなく，今回採用した基準値に大きな問題はないと考えられるため，本論文では比較手法 1 における基準値を 50%として検証・考察を進める．
[*10] 5.2 節で後述する検討用データセットに対して閾値を変更して精度を検証したところ，閾値を 1 件としたときが最も精度が良く，閾値を大きくすることで精度が低下していく傾向が見られた．そのため，今回採用した基準値に大きな問題はないと考え，本論文では比較手法 2 における基準値を 1 件以上として検証・考察を進める．

表 2 データセットの一部と留学生による分類結果
Table 2 A part of experiment data set and decision results by foreign students.

ID	語句	分類結果（2 名以上）	第 1 種の文化差	第 2 種の文化差	日本と中国の違いについて
30	火山	文化差なし			—
38	戸籍	文化差なし			—
57	1111 年	文化差なし			—
58	クロカジキ	日本発祥	○		標準和名は和歌山県田辺市周辺での呼び名に因む．
61	暦応	日本発祥	○		日本の年号．
72	水龍	日本発祥	○		大局将棋の駒の 1 つ．
80	日本アカデミー賞	日本発祥	○		日本のアカデミー賞．
113	ヒスイ	日本と中国は違う		○	軟玉は中国以外では半貴石に分類される．
124	ソンクラーン	日本と中国は違う		○	日本では水掛け祭りといういい方もすることがある．
130	雨女	日本発祥	○		日本の妖怪．
134	妃	日本と中国は違う		○	日本では天皇以外の男性皇族の配偶者に用いられる．
168	シラコバト	文化差なし			—
179	かごしま黒豚	日本と中国は違う		○	中国では「巴克夏猪」（バークシャーの意）と呼ぶ．
192	交通	日本と中国は違う		○	「日本の交通」という記事が存在する．
298	トシェビーチ	文化差なし			—
299	四酸化三コバルト	文化差なし			—

し」という 2 種類に分類した場合の判定精度を示す．なお，2 種類の分類においては，正解データおよび各手法による判定結果のうち，第 1 種または第 2 種の文化差である（と判定された）ものを「文化差あり」とし，判定精度を検証する．

5.3 評価用データセットを用いた検出実験

本節では，提案手法が，あらかじめ絞られた語句だけでなく，Wikipedia の記事全体に対しても有効かどうか検討する．本実験で使用するデータは，日本語版 Wikipedia の記事データは，2012 年 7 月 18 日，また中国語版 Wikipedia の記事データは，2012 年 7 月 28 日のダンプデータを使用している．

まず，文化差の検出精度を検証するため，評価用データセットを作成する．検討用データセットでは，人名や固有名詞は，ほとんど第 1 種の文化差であり，実験するのに適さない可能性があるため，データセットから除外した．しかし，評価用データセットにおいては，そういった記事についても何らかの傾向が見られる可能性を考慮して，Wikipedia 全体の記事を用いることとした．データセットの作成手順を以下に示す．

Step 1 Wikipedia の全記事データに対して，提案手法を適用する．

Step 2 検出結果から「第 1 種の文化差」「第 2 種の文化差」「文化差なし」となった語句をそれぞれ 100 件ずつランダムに抽出する．

Step 3 本学の中国人留学生 5 名に依頼し，次の 3 種類に分類してもらう．

- 中国のものと，日本のものは同じ（文化差なし）．
- 基本的には同じだが，発祥など少し違いがある（第 1 種の文化差）．
- 中国と日本のものは違う（第 2 種の文化差）．

「基本的には同じだが，発祥など少し違いがある」または，「中国と日本のものは違う」を選択した場合には，どのように違うかを簡単に記述してもらった．

表 2 に，利用したデータセットの一部と留学生による分類結果を示す．また，表に示した分類結果は，5 名の留学生による分類結果を集計し，基本的に多数決で判断した．しかし，第 1 種の文化差，または第 2 種の文化差に 1 票でも入っていた場合，文化差ありを優先し，そのどちらかを多数決で判断した．今回の分類は，4.1 節における分類と違い，ランダムに語句を収集しており，日本人であっても意味を知らなかったり，文化差があるかどうか判定が難しい語句が含まれていると考えられるため，1 人でも文化差を認識した場合，その意見を尊重するような分類規則とした．なお，作成したデータセットに含まれる 300 語句中，「第 1 種の文化差」がある語句は 164 件，「第 2 種の文化差」がある語句は 23 件，「文化差なし」である語句は 113 件である．

6. 実験結果と考察

6.1 検討用データセット

表 3 に，それぞれの手法の精度の比較を示す．結果として，すべての文化差について提案手法が最も高い精度を示した．また，**表 4** に各文化差における判定数と正解率を示す．

比較手法 1 と比較手法 2 は，記述量が多い記事をその国・言語のもの（第 1 種の文化差）と判断しやすい傾向がある．検討用データセットは，「日本独特のもの」「日本にも海外にもあるが，日本と海外とは違うもの」を収集したもので

表 3 検討用データセットを用いた各手法の精度の比較
Table 3 Comparison with each methods in accuracy using data set for examination.

	第 1 種の文化差			第 2 種の文化差			文化差あり			文化差なし		
	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値
比較手法 1	0.406	0.667	0.505	—	—	—	0.783	0.643	0.706	0.333	0.500	0.400
比較手法 2	0.667	0.762	0.711	0.516	0.390	0.444	0.797	0.750	0.773	0.400	0.467	0.431
比較手法 3	0.722	0.619	0.667	0.826	0.463	0.594	0.983	0.699	0.817	0.545	0.968	0.698
提案手法	0.787	0.881	0.831	0.788	0.634	0.702	0.900	0.878	0.889	0.706	0.774	0.738

表 4 検討用データセットを用いた各判断における判定数と正解率
Table 4 The number of judgments and accuracy rate of each judgment using data set for examination.

	第 1 種の文化差	第 2 種の文化差	文化差なし
A	—	—	—
B	2 (50%)	—	—
C	—	7 (71.4%)	—
D	—	0	—
E	0	—	—
F	—	3 (100%)	—
G	31 (87.1%)	—	—
H	1 (100%)	9 (88.9%)	—
I	—	—	7 (85.7%)
J	—	—	7 (71.4%)
B’	1 (100%)	—	—
C’	—	1 (100%)	3 (100%)
D’	—	2 (100%)	—
E’	1 (100%)	—	—
F’	—	2 (50%)	17 (64.7%)
G’	11 (54.5%)	—	—
H’	—	9 (66.7%)	—
平均	81.9%	82.4%	80.45%

※ダッシュ記号のある判断はフローチャートの右側のものである．

ある．現在，中国語版の Wikipedia に比べて，日本語版の Wikipedia の多くの記事の記述量は多いため，記事の多くが「日本発祥のものである」と判定されやすい．そのため，比較手法 1 および比較手法 2 の第 1 種の文化差の再現率は高くなると考えられる．しかし，両手法とも再現率はそれほど高い値を示さなかった．これは，今回選ばれた検討用データセットは人が思いついた語句を収集しており，語句として使用される頻度が高いので，それぞれの Wikipedia における項目数では差が見られなかったためであると考えられる．

また，比較手法 1 および比較手法 2 は，文化差なしの F 値が低くなると考えられる．これは，多くの語句が，「第 1 種」あるいは「第 2 種」と判断されると考えられるためである．しかし，両手法とも検討用データセットにおける実験では再現率はそれほど低い値ではない．これは，上述したように検討用データセットでは，項目数による差はあまり存在しなかったためであると考えられる．

比較手法 3 と比較した提案手法の特徴としては，次が考えられる．比較手法 3 は日本語版 Wikipedia のみを利用するが，提案手法は，日本語版 Wikipedia および中国語版 Wikipedia の両方を利用する．提案手法は，言及表現だけでなく，「日本の文化に関連した書きかけ項目です」や「世界的な観点からの説明がされていないおそれがあります」などの Wikipedia の構造的な特徴を利用している．そのため，言及表現だけで判定するには，不適切であった記事を適切に判定できていると考えられる．また，比較手法 3 は，日本語版には適用可能であるが，他の言語版に適用が困難である．提案手法は，カテゴリの分類および，各国への言及率を用いており，他の言語版への拡張が容易である．

また表 4 より，判断 A で分岐した後の判断を見比べると，判断 A の条件にあてはまらなかった場合に適用される判断（判断 B’～H’）は，あてはまった場合（判断 B～H）よりも判定精度が低くなっていることが分かる．これは，判断 A の条件にあてはまった場合，「文化差なし」については考慮しなくてもよいが，あてはまらなかった場合は「文化差なし」を考慮しなければならないことが影響していると考えられる．

6.2 評価用データセット

表 5 に，それぞれの手法の精度の比較を示す．**表 6** に，各文化差における判定数と正解率を示す．

表 3 と表 5 の提案手法について見ると，すべての文化差判断において，F 値が低下している．第 1 種の文化差については，適合率が向上したものの，再現率が低下した．第 2 種の文化差については，再現率が向上したものの，適合率が低下した．文化差あり，文化差なしについても，適合率，再現率，F 値が低くなっている．また表 5 より，他のすべての手法においても，第 2 種の文化差における F 値が低下している．表 6 を見ても，判断 C は判定数が 1 つで 100%であるものの，他のすべての判断において，正解率が 10%程度である．このことから，検討用データセットの分析を通して得られた指標を用いた提案手法は，評価用データセットにおける第 2 種の文化差の判定では，判断としての信頼性が低いことが考えられる．提案手法，比較手法 3 における第 2 種の文化差と判定する条件としては，記述内容の言及表現を用いた判断が多数を占めている．これは，検討用データセットの記述内容のみを分析しており，

表 5 評価用データセットを用いた各手法の精度の比較
Table 5 Comparison with each methods in accuracy using data set for evaluation.

	第 1 種の文化差			第 2 種の文化差			文化差あり			文化差なし		
	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値
比較手法 1	0.565	0.817	0.668	—	—	—	0.662	0.840	0.741	0.524	0.292	0.375
比較手法 2	0.846	0.701	0.767	0.136	0.348	0.195	0.754	0.786	0.770	0.314	0.292	0.303
比較手法 3	0.921	0.354	0.511	0.167	0.174	0.170	0.908	0.422	0.577	0.493	0.929	0.664
提案手法	0.920	0.561	0.697	0.150	0.652	0.244	0.740	0.791	0.765	0.610	0.540	0.573

表 6 評価用データセットを用いた各判断における判定数と正解率
Table 6 The number of judgments and accuracy rate of each judgment using data set for evaluation.

	第 1 種の文化差	第 2 種の文化差	文化差なし
A	—	—	—
B	0	—	—
C	—	1 (100%)	—
D	—	0	—
E	0	—	—
F	—	14 (14.3%)	—
G	43 (100%)	—	—
H	5 (60%)	9 (11.1%)	—
I	—	—	89 (62.9%)
J	—	—	4 (50%)
B’	0	—	—
C’	—	0	0
D’	—	9 (11.1%)	—
E’	0	—	—
F’	—	24 (16.7%)	7 (85.7%)
G’	52 (86.5%)	—	—
H’	—	43 (13.9%)	—
平均	82.1%	27.9%	66.2%

※ダッシュ記号のある判断はフローチャートの右側のものである．

今後，その他の語句の記述内容の特徴も考慮にいれる必要がある．

また，検討用データセットにおける評価者の判断は，あらかじめ日本人が選択した文化差があると感じられる語句を評価している．しかし，評価用データセットでは Wikipedia からランダムで選択された語句を評価している．評価者が，一般的でないが存在する第 2 種の文化差を「文化差なし」と判定している可能性がある．その理由として，語句を第 2 種の文化差と判定するには，両国に関する深い知識が必要となるため，第 2 種の文化差と評価するのが難しかった可能性が考えられる．また両国に深い知識を持っており，はっきり語句に文化差がないと思っていても，その語句には第 2 種の文化差があり，提案手法があまり知られていない文化差を可視化しているということも考えられる．

ほかにも，Wikipedia で検出される語句の国ごとの違いが，語句そのものの意味や概念の違いであるかどうか適切な判断をする必要がある．たとえば，第 2 種の文化差と判定された語句として，「投資信託」という語句がある．これは，日本語版 Wikipedia においては，日本について言及されており，中国版 Wikipedia においても言及されており，判断 F において文化差があると判定される．しかし，「投資信託」の根本的な意味や概念に違いはなく，「投資信託」周辺の細かな制度などの違いが存在するだけである．「投資信託」の根本的な意味や概念が，それぞれの国の人にとって同じで，認識の差などが存在しないことが考えられる．これが文化の差であるといえるか検討する必要がある．ほかにも，「政治」や「経済」という語句が存在する．これらは，ともに「日本の政治」「日本の経済」という記事が存在し，判断 D において第 2 種の文化差と判定される．しかし，それぞれの国の人がイメージする，「政治」と「経済」の意味や概念が大きく違うとは考えにくい．検討用データセットは，もともと文化差があると考えられる語句を集めており，このような違いがある語句が存在しなかったため，文化差の検出精度が高かったことが考えられる．今後は，国ごとの違いにおいて，どのような特徴があれば，違いの中でも「文化の違い」であるといえるのかを検討する必要がある．

また，提案手法における文化差判定時に，効果的であった指標について確認した．表 4 と表 6 の正解率を見ると，両実験において，判断 C（世界的な観点から説明されていない記事），判断 G と G’（定義分とカテゴリにおける国名数と言及表現の有無），判断 I（各言語版 Wikipedia 内に言及表現がない），そして判断 F’（日本語版 Wikipedia で日本に言及しているが，中国語版 Wikipedia で「日本」という表記が存在しない場合）における第 2 種の文化差の検出条件が比較的高い文化差検出精度を示した．これらは，文化差判定時に，有効な指標であると考えられる．

6.3 他文化圏への応用

今回提案した手法は，日中間における文化差判定を考慮した手法であるが，今後，他文化圏において文化差を判定することを想定しており，日中間に特化した判断は用いていない．したがって，本提案手法はそのまま他文化圏における文化差判定においても適用自体は可能である．ただし，他文化圏においても同程度の検出精度を得られるかどうかは今後検証する必要がある．

7. おわりに

今回，多言語知識のデータベースとして Wikipedia を利用し，文化差を検出する手法を提案した．本論文の貢献は，以下の 2 点にまとめられる．

- 文化差判定時に指標となる特徴の抽出
  文化差判定時の指標における特定の語句に関する特徴として，本文中の特徴，カテゴリ中の特徴，タイトル中の特徴があり，すべての語句に関する特徴として，項目数の特徴，国名数・言語名数の特徴，執筆者の意図の特徴が存在することを明らかにした．
- 文化差検出手法の提案
  提案手法は，検討用データセット 114 語句を用いた実験において，高い精度を示した．また，文化差判定時に効果的であると考えられる判断として
  – 判断 C：世界的な観点から説明されていない記事
  – 判断 F：日本語版 Wikipedia で日本に言及しているが，中国語版 Wikipedia で「日本」という表記が存在しない場合
  – 判断 G：定義文とカテゴリにおける国名数と言及表現の有無
  – 判断 I：各言語版 Wikipedia 内に言及表現がない

  の 4 つの判断が存在することが分かった．

今後の課題としては，あらかじめ語句の文化差を評価するのでなく，手法による文化差判定結果についての評価も検討する必要がある．そして，提案手法のさらなる文化差判定精度の向上を検討する．文化差検出精度の向上を考慮するにあたり，文化差検出精度の良かった判断 C，G，G’，I，F’ の判断に対して重みを考慮した手法が考えられる．また，異なる文化圏との文化差の検出においても適用可能かどうか検討する必要がある．

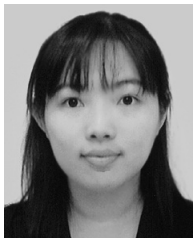
謝辞 本研究の一部は，独立行政法人科学技術振興機構研究成果 最適展開支援事業（A-STEP）探索タイプおよび和歌山大学平成 25 年度独創的研究支援プロジェクトの補助を受けた．

参考文献

[1] 藤井薫和，重信智宏，吉野　孝：機械翻訳を用いた異文化空間チャットコミュニケーションにおけるアノテーションの評価，情報処理学会論文誌，Vol.48, No.1, pp.63–71 (2007).
[2] 藤井薫和，吉野　孝：異文化間コミュニケーション支援のためのアノテーション自動獲得システムの開発，情報処理学会研究報告，グループウェアとネットワークサービス研究会，2008-GN-66, pp.141–146 (2008).
[3] 岡本健吾，吉野　孝：会話中の名詞の関連情報を用いた対面型異文化間コミュニケーション支援システムの構築と評価，情報処理学会論文誌，Vol.52, No.3, pp.1213–1223 (2011).
[4] Cho Heeryon，石田　亨，山下直美ほか：絵文字解釈における人間の文化差判定，ヒューマンインタフェース学会論文誌，Vol.10, No.4, pp.427–434 (2008).
[5] Koda, T. and Ishida, T.: Cross-cultural Study of Avatar Expression Interpretations, *SAINT 2006*, pp.130–136 (2006).
[6] Hofstede, G.: *Cultures and Organizations: Software of the Mind*, McGraw-Hill, London (1991).
[7] Pfeil, U., Zaphiris, P. and Ang, C.A.: Cultural Differences in Collaborative Authoring of Wikipedia, *Journal of Computer-Mediated Communication*, Vol.12, pp.88–113 (2006).
[8] 藤原裕也，灘本明代：Wikipedia の言語間比較による差異情報抽出手法の提案，情報処理学会研究報告，Vol.2011-DBS-152, No.3, pp.1–8 (2011).
[9] 松浦愛美，江口浩二：時系列対訳トピックモデルを用いた言語横断トレンド分析，情報処理学会研究報告，Vol.2010-DBS-75, No.11, pp.1–5 (2010).
[10] 吉岡真治：Wikipedia を用いた中日カタカナ翻訳辞書の作成と言語グリッドへの応用，電子情報通信学会技術報告，人工知能と知識処理，Vol.109, No.424, pp.43–46 (2010).
[11] 西田ひろ子：異文化間コミュニケーション，創元社 (2000).

諏訪 智大 （学生会員）

1991 年生．2013 年和歌山大学システム工学部デザイン情報学科卒業．現在，同大学大学院システム工学研究科システム工学専攻博士前期課程在学中．文化差検出に関する研究に従事．



宮部 真衣 （正会員）

1984 年生．2006 年和歌山大学システム工学部デザイン情報学科中退．2008 年同大学大学院システム工学研究科システム工学専攻博士前期課程修了．2011 年同大学院システム工学研究科システム工学専攻博士後期課程修了．博士（工学）．現在，京都大学学際融合教育研究推進センターデザイン学ユニット特定研究員．コミュニケーション支援に関する研究に従事．

吉野 孝 （正会員）

1969 年生．1992 年鹿児島大学工学部電子工学科卒業．1994 年同大学大学院工学研究科電気工学専攻修士課程修了．博士（情報科学）．現在，和歌山大学システム工学部教授．