

Wikipedia の言語間差異情報を用いた文化差検出手法の提案

Proposal of Cultural Difference Detection Method Using Difference Information of Wikipedia

諏訪 智大[†] 宮部 真衣[‡] 吉野 孝^{††}
Tomohiro Suwa Mai Miyabe Takashi Yoshino

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、画像等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。

我々はこれまでにいくつかの文化差検出手法を提案してきた [4]。しかし、これらの手法の検証実験を通じて、従来の文化差の定義では、捉えきれない文化差が存在する可能性が示唆された。そこで本稿では、新たに文化差を定義しなおし、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。

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 [5]。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、アバター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。アバター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 [6]。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。Ulrike 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 [7] 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関連することを明らかにした [8]。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 [9]。

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 [10]。例えば、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 [10]。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。そこで本稿では、特に「知識」の面から「文化」を捉えることとし、まず、文化差検出手法の第一歩として、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、文化差の定義について説明する。本研究では、多言語間リンクでつながっている、同じ語句を説明する Wikipedia の言語間差異情報を双方の国における、認識の差・文化の差として捉えることとする。まず、各言語版の Wikipedia にどれほどの差異情報が存在するかを、集合論のベン図で表す。次に、集合の大きさをその国における Wikipedia の記述量の多さとして、集合の重なり具合により言語間の共通情報・差異情報を表現する。考えられる文化差パターンを図 1 に示す。また、それぞれの集合についての説明も示す。

- (a) 文化差なし
情報の差がなく、文化差がないと考えられる。
- (b) 部分集合の文化差
大きい集合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、大きい方の集合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と考えられる。
- (c) 一部が同一の文化差
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と考えられる。
- (d) 全く異なる文化差
両文化圏において、全く異なる概念であると考えられる。

4. Wikipedia の言語間差異情報の抽出手法

本稿では、言語間差異情報として、両言語版における名詞の重なり率を扱うこととする。比較する両言語版 Wikipedia から名詞のみを抽出し、比較を行う。以下に詳しい手順を示す。

- (1) 語句と対応する日本語版 Wikipedia 記事を特定
- (2) 抽出した日本語版 Wikipedia の多言語間リンクから中国語版 Wikipedia 記事を特定
- (3) 日本語版 Wikipedia の記事を形態素解析^{*1}にかけ、名詞のみを抽出し、重複した名詞を削除
- (4) 比較する言語版 Wikipedia の記事を日本語に翻訳^{*2}し、形態素解析にかけ、名詞のみを抽出し、重複した名詞を削除

^{*1} MeCab を用いた

^{*2} Google 翻訳を用いた

[†] 和歌山大学大学院システム工学研究科, Graduate School of Systems Engineering, Wakayama University

[‡] 京都大学デザイン学ユニット, Unit of Design, Kyoto University

^{††} 和歌山大学システム工学部, Faculty of Systems Engineering, Wakayama University

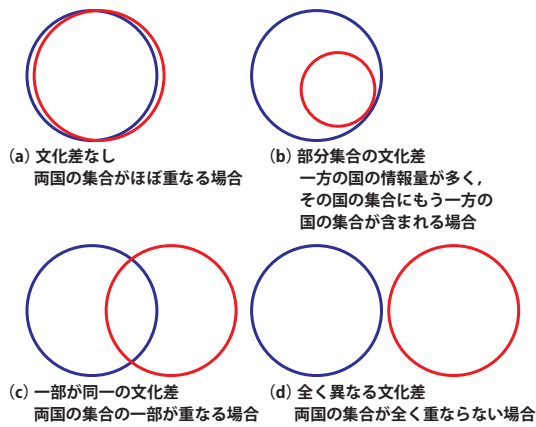


図 1: 文化差パターン

(5) 両言語版から抽出した名詞のうち、Wikipedia 記事として存在しない語句を削除

(6) 最終的に残った両言語版の名詞リスト同士を比較

Wikipedia 記事として存在しない語句を削除した理由は、Wikipedia に記事として存在しない語句は、文化差検出において、本質とは関係がない語句である可能性を考慮したためである。

5. 評価実験の結果と考察

4 章で述べた手法を用いて、語句における集合の重なり率の傾向を調査する。今回、日本と中国における語句の文化差について調査を行った、用いた語句は、従来の文化差検出手法 [4] の検討において用いられた語句である。

実験の結果から得られた、何らかのパターンが見られた語句を示す。表 1 は文化差パターン (b)、表 2 は文化差パターン (c) の傾向を示した語句を示す。表はそれぞれ、日中で比較した語句数と、それぞれの国から見た語句の重なり率を示している。

表 1 において示した語句は、文化差パターン (b) の傾向が見られた。日本語版 Wikipedia における名詞数と中国語版 Wikipedia における名詞数を比較して、大幅な情報量の差が存在するという傾向が見られた。小さい方の集合が大きい方の集合の部分集合とはならなかった。重なり率としては、中国側から見て 50% 程度であり、日本側から見ると低い値となったことから、伝わっている知識が限定的であると言える。またそれぞれの語句について見ると、いずれの語句についても日本について深く関係のある語句であることがわかる。

表 2 において示した語句は、文化差パターン (c) の傾向を明確に示した。それぞれの情報量はほぼ同じで、語句の重なり率が低く、それぞれの国の Wikipedia において独自の説明を記述している可能性がある。またそれぞれの語句についての Wikipedia 記事を見ると、これらの語句は、それぞれの文化圏で意味の異なることを記述していることがわかった。例えば「元日」の中国語版 Wikipedia を見ると、日本との旧正月の違いを説明している。また「煎餅」は、日本ではお菓子であるが、中国では主食として食べられている。他の語句においても「文化差あり」に分類されている [4]。

今回の実験では、文化差パターン (a) と文化差パターン (d) の傾向を示す語句は見られなかった。文化差パターン (a) が見られなかった理由としては、両言語版の Wikipedia において、語句に文化差がなかったとしても、同じ語句を用いて、かつ同じような説明をしていることが少ない可能性が考えられる。また、翻訳処理によって、本来ならば同じ語句であっても、異なる表現になっている可能性も考えられる。文化差パターン (d) については、そもそも Wikipedia としては異なる語句のため多言語間リンクが存在していない可能性が考えられる。

表 1: 文化差パターン (b) の傾向を示した語句

	日	中	重なり数	重なり率 (日)	重なり率 (中)
パチンコ	866	49	29	3.3%	59.2%
神道	411	71	39	9.5%	54.9%
節分	359	20	10	2.8%	50.0%
相撲	710	128	64	9.0%	50.0%
おにぎり	417	39	20	4.8%	51.3%

表 2: 文化差パターン (c) の傾向を示した語句

	日	中	重なり数	重なり率 (日)	重なり率 (中)
元日	117	78	12	10.3%	15.4%
煎餅	256	320	44	13.8%	17.2%
老人 (高齢者)	183	166	32	17.5%	19.3%
公衆便所	162	124	28	17.3%	22.6%
クリスマス	436	421	126	28.9%	29.9%

6. 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。本稿の貢献は、以下にまとめられる。

- 文化差検出手法の提案
新たな文化差定義に対応した、文化差検出手法を提案した。
- 文化差検出評価
提案手法は、文化差パターン (b) (c) において定義に近い傾向が見られた。

今後の課題としては、今回は Wikipedia に記事として存在する語句のみを比較したが、語句として比較すべき語句とそうでない語句を適切に選別することをさらに検討する必要がある。また、語句の比較方法として完全一致ではなく、語句の意味の近さを考慮に入れた比較方法も考えられる。そして提案手法が、より多くの語句で文化差定義のパターンを明確に示すようにする必要がある。

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「Wikipedia と機械翻訳を用いた多言語間における文化差検出サービス」の補助を受けた。

参考文献

- 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)。
- 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)。
- 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)。
- 諏訪智大, 宮部真衣, 吉野 孝: Wikipedia を用いた文化差検出における誤検出語句の分析, 電子情報通信学会技術報告, 人工知能と知識処理研究会, Vol.112, No.435, pp.11-16 (2013)。
- Cho Heeryon, 石田 享, 山下直美, 稲葉利江子, 高崎俊之, 神田智子: 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)。
- Tomoko Koda and Toru Ishida: Cross-cultural study of avatar expression interpretations, SAINT 2006, pp.130-136 (2006)。
- Hofstede G: Cultures and Organizations: Software of the Mind. London: McGraw-Hill (1991)。
- Ulrike Pfeil, Panayiotis Zaphiris, Chee Siang Ang: Cultural Differences in Collaborative Authoring of Wikipedia, Journal of Computer-Mediated Communication, 12, pp.88-113 (2006)。
- 藤原裕也, 瀬本明代: 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)。
- 西田ひろ子: 異文化間コミュニケーション, 創元社 (2000)。