E-044

Wikipedia を用いた文化差検出手法の評価
Evaluation of Cultural Difference Detection Method using Wikipedia

吉野　孝 †
Takashi Yoshino
宮部真衣 ‡
Mai Miyabe

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]．我々はこれまでに，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，誤解を減らす工夫を行った [2, 3]．しかし，アノテーションの付与が必要となる文化差がある（と考えられる）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であるため，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．

そこで我々は，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．本稿では，Wikipedia を用いて，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文化差の検出手法の評価について報告する．

2. 関連研究

Cho らは，絵文字が異文化間で普遍的に解釈されない問題に着目し，解釈に文化差のある絵文字の検出おける工学的な手法の適用可能性について検討し，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 [4]．Koda らは，アバターの表情に関する異文化間の解釈について実験し，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．

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，Wikipeida を用いた [6]．吉岡は，機械翻訳システムの辞書構築のために，Wikipedia の言語間リンクを用いて中日の翻訳辞書を作成する方法を提案している [7]．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. 提案手法

3.1 文化差の定義

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，一概に定義することは困難である [8]．例えば欧米で用いられる「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 [8].「文化」を単純に「測る」ことは困難であるが，本稿では，特に「知識」の面から「文化」を捉えることとし，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」のある語句は，一方の文化圏で発生し，相手の文化圏に存在しない，あるいは伝わっていないものであるとする．例えば,「モスバーガー」は日本発祥のため，日本語の説明は存在するが，他の言語の説明はほとんど存在しない.「第 2 種の文化差」のある語句は，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とする．例えば,「醤油」は日本と中国のどちらにも存在するが，お互いに意味が異なる．

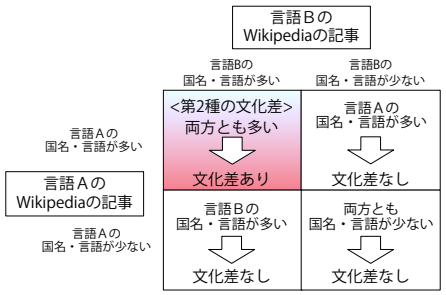

図 1: 第 2 種の文化差の検出手法（国名・言語の数を利用）

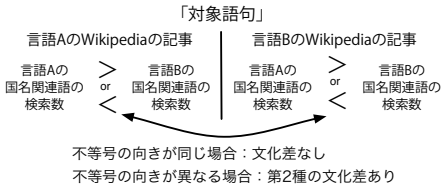

図 2: 国名関連語の検索数を利用した第 2 種の文化差の判定処理

3.2 Wikipedia を用いた文化差の検出方法

第 1 種の文化差の検出は，Wikipedia における記事の言語間リンクの有無を利用して可能であると考えた．どちらか一方のみに記事があれば,「文化差がある」（＝第 1 種の文化差）と判定する§.

第 2 種の文化差の検出は，Wikipedia の記事に含まれる国名・言語の数を利用して可能であると考えた．図 1 に第 2 種の文化差の検出方法を示す．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・言語が多い場合には，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」（＝第 2 種の文化差）と判定する．

本稿では，第 2 種の文化差検出手法の評価について検討を進める．

†和歌山大学システム工学部
‡東京大学知の構造化センター

§単純に，記事が作成されていないだけの場合も考えられる．

表 1: 日本語版 Wikipedia と他言語版 Wikipedia 間のカテゴリ別の第 2 種の文化差の検出数

カテゴリ	記事の件数	中国	韓国	米国・英国	ポルトガル	フランス	ドイツ	スペイン	平均	変動係数
人物記	205	2 %(4)	1 %(2)	3 %(6)	1 %(3)	4 %(9)	1 %(3)	1 %(2)	2 %(4.1)	0.6
哲学と心理学	17	0 %(0)	0 %(0)	12 %(2)	0 %(0)	6 %(1)	0 %(0)	0 %(0)	3 %(0.4)	1.7
宗教	25	12 %(3)	**20 %(5)**	16 %(4)	8 %(2)	8 %(2)	8 %(2)	4 %(1)	11 %(2.7)	0.5
人文科学	70	6 %(4)	7 %(5)	19 %(13)	7 %(5)	**27 %(19)**	**23 %(16)**	13 %(9)	14 %(10.1)	0.5
言語と文学	46	4 %(2)	4 %(2)	13 %(6)	4 %(2)	17 %(8)	15 %(7)	9 %(4)	10 %(4.4)	0.5
科学	299	2 %(6)	0 %(1)	18 %(55)	2 %(5)	11 %(33)	8 %(24)	4 %(12)	6 %(19.4)	0.9
技術	71	4 %(3)	6 %(4)	14 %(10)	3 %(2)	11 %(8)	11 %(8)	4 %(3)	8 %(5.4)	0.5
芸術と娯楽	76	0 %(0)	1 %(1)	**21 %(16)**	7 %(5)	**24 %(18)**	**20 %(15)**	5 %(4)	11 %(8.4)	0.8
歴史と地理	191	8 %(16)	16 %(30)	17 %(32)	6 %(12)	18 %(34)	15 %(28)	10 %(19)	13 %(24.4)	0.3
合計	1000	4 %(38)	5 %(50)	14 %(144)	4 %(36)	13 %(132)	10 %(103)	5 %(54)	8 %(79.6)	0.5

※ () 内は件数を示す．変動係数は標準偏差の値を平均で割った値である．

3.3 文化差の検出手順

第 2 種の文化差の検出手順を次に示す．

(1) 国名・言語の検索処理
国名・言語の検索処理では，記事内の国名・言語を検索する．検索する語句は,「国名」「国名の」「言語」「言語の」「〜人」「〜人の」を意味する語句である．例えば,「日本」に関する記事を検索する場合，各言語版の Wikipedia の記事において,「日本」「日本の」「日本語」「日本語の」「日本人」「日本人の」を意味する各言語での語句（例えば英語では「Japan」「Japanese」）の検索を行う．本稿では以降，上記の検索語句を,「国名関連語」と示す．

(2) 文化差の判定処理
文化差の判定処理では，各記事内に含まれる，国名関連語の数を比較する．図 2 に，国名関連語の数を利用した第 2 種の文化差の判定処理を示す．各言語の Wikipedia の記事に含まれる国名関連語の数を比較し，不等号の向きが同じ場合には，文化差なしと判定する．不等号の向きが異なる場合には，文化差ありと判定する．

今回提案した手法では，図 1 で提案している第 2 種の文化差の判定における「国名・言語が**多い**」という判定条件において,「多い」と判定する値を 10 件と暫定的に決めた．検索数の差異の判断については，1 つ以上違えば文化差があったと判定している¶．

4. 実験方法

本研究では，Wikipedia を多言語知識のデータベースとして用いている．今回，文化差検出の基準となる言語は日本語版（72.7 万件）とした．比較対象の言語としては，Wikipedia 上において登録記事数の多い，英語版（352.5 万件），ドイツ語版（117.4 万件），フランス語版（105.4 万件），スペイン語版（69.9 万件），ポルトガル語版（66.6 万件）の 5 言語および中国語版（34.0 万件），韓国語版（15.3 万件）を含めた，合計 7 言語を用いることとした‖．

Wikipedia は，言語によって登録記事数に大きな差がある．そこで，文化差の検出効果を検証する分野（記事）として,「すべての言語版にあるべき項目の一覧**」（以下,「Wikipedia 作成優先項目」と略す）を用いた††．「Wikipedia 作成優先項目」は，全ての言語版が最低限の有用な内容を持つように促すことを目的に選択されており，世界的な著名人や宗教，国名，社会問題，言語，科学などの幅広い分野にわたる 1000 項目が列挙されている．

¶今後，検索数の閾値や検索数の違いの判断基準については，検討が必要である．

‖Wikipedia 上の登録件数は 2011 年 1 月 10 日現在．

**http://ja.wikipedia.org/wiki/Wikipedia:すべての言語版にあるべき項目の一覧

††記事は，Web 上から 2011 年 1 月 10 日に取得．

5. 実験結果

表 1 に，日本語版 Wikipedia と他言語版 Wikipedia 間のカテゴリ別の第 2 種の文化差の検出数を示す．カテゴリは，Wikipedia 上の分類を利用した．最も文化差の検出数が多い言語が英語版（14 %，144 件）であり，もっとも少ない言語がポルトガル語版（4 %，36 件）である．

今回，検出割合が 20%以上の場合，検出数が多いと見なす．文化差の検出割合が 20%を越えたのは，表 1 の太字下線の 6 箇所である．カテゴリ別の平均を見ると,「人物記」「哲学と心理学」「科学」「技術」において，文化差の検出数は少ない．特に,「人物記」「哲学と心理学」は，平均で，2%〜3%である．

6. おわりに

今回，多言語知識のデータベースとして Wikipedia を用い，記事に含まれる国名や言語（国名関連語）の数によって，第 2 種の文化差検出手法の評価を行った．

本稿の貢献は次の 2 点にまとめられる．

(1) Wikipedia を用いた第 2 種の文化差の検出結果をカテゴリ別に見た場合，カテゴリによって差が見られることを示した．

(2) カテゴリ別の平均を見ると,「人物記」「哲学と心理学」「科学」「技術」において，文化差の検出数は少なく，特に,「人物記」「哲学と心理学」は，平均で，2%〜3%であることを示した．

今後の課題として，第 2 種の文化差の検出精度の検証がある．また，記事に含まれる国名や言語（国名関連語）の検索数が近接している場合の判定基準の検討が必要である．最終的には，他の多言語サービスへ提供できるように，Web サービス化を行う．

謝辞

本研究の一部は，日本学術振興会科学研究費 基盤研究 (B)(22300044) 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] 藤井薫和ほか：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，情処研報，2008-GN-66，pp.141-146 (2008).
[3] 岡本健吾ほか：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，情処学論，Vol.52，No.3，pp. 1213-1223 (2011).
[4] Cho Heeryon 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，Vol.10，No.4，pp.427–434 (2008).
[5] Tomoko Koda et. al: Cross-cultural study of avatar expression interpretations, SAINT 2006, pp.130–136 (2006).
[6] 松浦愛美ほか：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，情処研報，2010-DD-75，No.11，pp.1–5 (2010).
[7] 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，信学技報，人工知能と知識処理，Vol.109，No.424，pp.43–46 (2010).
[8] 西田ひろ子：異文化間コミュニケーション，創元社 (2000).