

Wikipedia を用いた文化差判定の検討

吉野 孝[†][†] 和歌山大学システム工学部宮部 真衣[‡][‡] 和歌山大学大学院システム工学研究科

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り、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。

そこで我々は、文化差を検出する仕組みとして、多言語知識のデータベースである、Wikipedia の利用による文化差の検出手法について検討を行った。本稿では、Wikipedia を用いた文化差判定に関する検討内容について報告する。

2 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、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 [4]。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 [4]。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量の違いで文化を測ることとした。

3 提案方式

本研究では、Wikipedia を多言語知識のデータベースとして用いることにした。文化差検出の基準となる言語は日本語 (72.7 万件) とした。Wikipedia での比較対象の言語としては、Wikipedia 上において登録数の多い、英語 (352.5 万件)、ドイツ語 (117.4 万件)、フランス語 (105.4 万件)、スペイン語 (69.9 万件)、ポルトガル語 (66.6 万件) の 5 言語および中国語 (34.0 万件)、韓国語 (15.3 万件) を含めた、合計 7 言語を用いることとした¹。

次の比較により、Wikipedia を用いた文化差判定が可能かどうかについて検討を行った。

(1) Wikipedia の項目 (多言語リンク) の有無の比較
日本語の用語からの他の用語への多言語リンクの有無を調べた。リンクがない場合には文化差があると判定する。

(2) Wikipedia の項目数の比較

それぞれの用語における Wikipedia の項目数を比較した。項目数の計数には、「< span class="toctext">」を用い、大項目や小項目などの項目のレベルは同一に取り扱った。比較は、2 段階で行った。

- (a) 日本語の用語にリンクがあり、多言語リンクがない場合は文化差があると判定する。
- (b) 日本語の用語および多言語リンクの両方がある場合には、項目数の違いを比較する。今回の検討では、項目数に 50 % 以上の違いがある場合に文化差があると判定する。なお、記事の量が少ない場合には、「< span class="toctext">」がないため、比較する際に、分母が 0 になる。そこで、全ての記事の概要部分を 1 項目とみなして、項目数の計数を行った。

上記の比較方法による効果を検証する分野として、文化差があると考えられる分野と文化差がない (少ない) と考えられる分野、文化差の有無が個別の項目に依存すると考えられる分野を用いた。それぞれの検証分野における利用データを以下に示す。検証分野のデータは、Web 上から 2011 年 1 月 10 日に取得したものを利用した。

● 文化差が ある と考えられる分野

- 日本観光に関する用語 (日本の観光地)
日本政府観光局 (JNTO) が行った訪問地満足度調査データ提供サービス²を利用した。地点数は、211 地点である。

● 文化差が ない (少ない) と考えられる分野

- コンピュータ用語
Wikipedia における「コンピュータ用語一覧³」における、日本語の項目に存在する 457 項目について評価を行った。
- Wikipedia 上の作成優先順位の高い項目 (Wikipedia 優先項目)
Wikipedia では、記事数の少ない言語において、まずどの記事を作成すべきかの指標となる項目として、「すべての言語版にあるべき項目の一覧⁴」がある。例えば、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項目が列挙さ

A Study on Cultural Difference Detection using Wikipedia

Takashi YOSHINO[†] Mai MIYABE[‡]

[†]Faculty of Systems Engineering, Wakayama University

[‡]Graduate School of Systems Engineering, Wakayama University

¹ Wikipedia 上の登録件数は 2011 年 1 月 10 日現在の値である。

² http://www.jnto.go.jp/jpn/tourism_data/data_service.html

³ <http://ja.wikipedia.org/wiki/コンピュータ用語一覧>

⁴ <http://ja.wikipedia.org/wiki/Wikipedia:すべての言語版にあるべき項目の一覧>

表 1: 比較項目の検索語句数と各言語のリンクの件数

	検索語句数	日本語	英語	中国語	韓国語	ドイツ語	フランス語	スペイン語	ポルトガル語
日本の観光地	211	205	156	160	137	144	137	118	115
コンピュータ用語	457	452	431	360	347	401	385	377	365
Wikipedia 優先項目	1000	1000	1000	996	998	994	996	997	998
動物名	437	311	305	237	118	247	279	228	186
世界の料理名	853	517	464	258	215	347	332	380	212

表 2: リンクの有無を利用した文化差判定結果

	英語	中国語	韓国語	ドイツ語	フランス語	スペイン語	ポルトガル語
日本の観光地	23.9%	22.0%	33.2%	29.8%	33.2%	42.4%	43.9%
コンピュータ用語	4.6%	20.4%	23.2%	11.3%	14.8%	16.6%	19.2%
Wikipedia 優先項目	0.0%	0.4%	0.2%	0.6%	0.4%	0.3%	0.2%
動物名	1.9%	23.8%	62.1%	20.6%	10.3%	26.7%	40.2%
世界の料理名	10.3%	50.1%	58.4%	32.9%	35.8%	26.5%	59.0%

表 3: 項目数を利用した文化差判定結果

	英語	中国語	韓国語	ドイツ語	フランス語	スペイン語	ポルトガル語
日本の観光地	85.9%	84.4%	94.1%	92.7%	94.6%	96.6%	96.6%
コンピュータ用語	44.0%	58.4%	72.8%	48.2%	47.3%	50.7%	63.3%
Wikipedia 優先項目	38.2%	44.1%	55.1%	36.6%	39.3%	38.3%	42.5%
動物名	46.0%	65.3%	95.8%	36.3%	43.4%	61.4%	87.1%
世界の料理名	54.4%	76.6%	79.1%	63.6%	64.8%	61.9%	82.2%

れている。この項目に含まれる 1000 項目について評価を行った。

- 文化差の有無が個別の項目に依存すると考えられる項目

- 動物名
上野動物園における飼育動物名一覧¹から、動物名を取得した。この項目に含まれる 437 項目について評価を行った。
- 世界の料理名
世界の料理がわかる辞典²から料理名を取得した。この項目に含まれる 853 項目について評価を行った。

4 実験結果と考察

2011 年 1 月 10 日時点の Wikipedia のデータを利用した。表 1 に、比較項目の検索語句数と各言語のリンクの件数を示す。「Wikipedia 優先項目」と「コンピュータ用語」以外は、Wikipedia と関係のないデータを利用しているが、多くの項目にリンクがあることが分かる。

表 2 に、リンクの有無を利用した文化差判定結果を示す。表中の割合は、日本語の用語にリンクがあり多言語リンクがない、文化差があると判定された項目の割合である。例えば、表 2 の「日本の観光地」の英語の割合は、表 1 から、 $1 - 156/205 = 0.239$ と求められる。「Wikipedia 優先項目」は文化差のある語句が 1% 以下、「コンピュータ用語」は 25% 以下であり、文化差のある語が少なくと判定されている。「日本の観光地」は比較的文化差があると考えられるが、英語では 23.9% の項目に、ポルトガル語で 43.9% の項目に文化差があると判定されている。「動物名」と「世界の料理名」は、言語によって大きく異なる判定結果がでている。

表 3 に、項目数を利用した文化差判定結果を示す。表中の割合は、項目数に 50% 以上の違いがあり、文化差があると判定された項目の割合である。文化差があると想定した「日本の観光地」は、80% 以上の項目で文化差があると判定された。ただし、文化差がないと想定した「Wikipedia 優先項目」と「コンピュータ用語」でも、半分程度の項目に文化差があると判定された。

¹ <http://zoollife.web.fc2.com/zoo/kantou/ueno/photo/z-ueno-p.htm>

² <http://kotobank.jp/dictionary/worldcook/>

項目数を利用した文化差判定は、リンクの有無だけを利用するよりも判定精度がよいと思われる。ただし、今回の検討では「項目数が 50 % 以上違う」項目に文化差があると判定したが、この判定基準は暫定的に決めたものであり、今後、閾値についての検討が必要である。また、「動物名」や「世界の料理名」などは、項目によって文化差が出ていると思われるため、各項目に関する詳細な分析が必要である。

5 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、リンクの有無や項目数の違いによって、文化差の判定が可能かどうかについて検討を行った。

本稿の貢献は次の 2 点にまとめられる。

- (1) Wikipedia を用いた文化差判定について、知識の量の違いによる判定手法を提案した。
- (2) リンクの有無、項目数の比較による文化差判定手法の評価を行った。項目数の比較を行った場合、日本の観光地に関しては、80% 以上の項目について、文化差があると判定された。また、「Wikipedia 優先項目」「コンピュータ用語」では、半分程度の項目に文化差があると判定された。

今後は、各評価項目の内容について詳細に分析を進める。

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, 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 (2011) .
- [4] 西田ひろ子: 異文化間コミュニケーション, 創元社 (2000) .