2C-7

文化差データの収集サービスの提案

吉野　孝 †　　宮部　真衣 ‡
† 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

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]. 文化の違いに基づく誤解を回避するためには，用いている単語に文化差があることを，話者に認識させる必要がある．しかし，相手の文化に関する十分な知識が必要となるため，文化差の有無の判断は容易ではない.

我々はこれまでに，文化差理解支援における，説明文および画像による文化差の可視化の効果の検証を行った．検証実験の結果，画像による文化差の可視化は，説明文に比べ，短時間での文化差の判定を可能にし，文化差判定の主観的な負荷は低いことを示した．しかし，現時点の「説明文」の提示および「画像」の提示は，いずれも文化差理解支援において，十分に貢献しない可能性があることもわかった [2].

そこで，低負荷かつ的確な文化差理解の支援手法を実現するためのリソースとして，Web 上に存在する「文化差の理解を促す事例」に着目した.「文化差の理解を促す事例」とは，誰かの経験などをもとに作成された，文化差が明確に記述された文章である.「文化差の理解を促す事例」は，文化の違いを端的に説明しており，これらを収集しておくことで，文化差理解支援における重要なリソースとして利用できる可能性がある．異文化間における誤解などの事例は，従来から様々な研究者が紹介している [3]．しかし，多様な語句に対する対応は困難であり，情報技術を用いて収集する必要がある.

本稿では，文化差データの収集サービスを提案し，サービスの実現に必要となる文化差データの収集手法および収集した文化差データの適切さについて述べる.

2　文化差データの収集サービス

2.1　文化差判定支援において提供すべき文化差データの要件

我々のこれまでの実験結果から,「説明文」[*1] の提示および「画像」[*2] の提示は，文化差の理解に関して，十分に貢献しない可能性のあることがわかった [2]．そこで，これまでの検証実験をもとに検討した，文化差判定支援において提供すべき文化差データの要件を以下に示す.

(1) **機械翻訳を利用せず，母語の情報を提示**
文献 [2] では，機械翻訳を用いて他国の語句の説明を行った．しかし，ぎこちない機械翻訳結果の解釈に大きな負担があり，十分に文化差の理解を促すことができないことがわかった．そのため，母語を用いた情報の提示が必要である.

(2) **具体的な違いの事例を提示**
文献 [2] では，語句の説明として，Wikipedia の概要を提示した．しかし，同じ語句を説明する各言語版 Wikipedia の概要の大部分は類似しており，文化差判定が容易でないことがわかった．そのため，具体的な違いに関する説明を提示する必要がある.

2.2　文化差データの収集手法

我々は，誰かの経験などをもとに作成された「文化差の理解を促す事例」は，2.1 節で述べた要件を満たすと考えた．そこで,「文化差の理解を促す事例」（以降，**文化差データ**と呼ぶ）の収集手法について検討を行った.

本研究では，文化差データの収集手法として，次の Web 検索エンジンを用いた手法を提案する.

(1) 「日本では」「中国では」＜検索語＞の 3 語を，完全一致検索を用いて，Web 検索を行う[*3].
(2) 検索結果に記載された Web ページを直接参照し，そのページに含まれるテキストデータを取得する.
(3) テキストデータから,「日本では」「中国では」「日本の＜検索語＞」「中国の＜検索語＞」のいずれかが含まれる文を,「文化差データの候補」とする.
(4) 「文化差データの候補」全体の中に，＜検索語＞が 1 個以上含まれている場合には,「文化差データの候補」全体を，その Web ページにおける＜検索語＞の「文化差データ」として蓄積する.

今回，検索語として,「日本では」「中国では」を用いた．これらを用いた理由は，執筆者が文化差を意図的に記述する際に，これらの語句が用いられる可能性が高いと考えたからである [4].

3　収集結果と考察

著者らが独自に作成した語句群（114 語）[*4] を検索語とし，文化差データの収集実験を行った．表 1 に「文化差あり」に分類された語句（83 語）を，表 2 に「文化差なし」に分類された語句（31 語）を示す．本実験では，Web 検索エンジンの上位 20 件を利用して文化差データを収集した[*5].

Proposal of Cultural Difference Data Collection Service
Takashi YOSHINO† Mai MIYABE‡
†Faculty of Systems Engineering, Wakayama University
‡Center for Knowledge Structuring, The University of Tokyo

[*1] 説明文の提示では，日本語と中国語の Wikipedia の説明文を用いた．中国語は，機械翻訳を用いて日本語にした.
[*2] 画像の提示では，Google 画像検索の検索結果を用いた.
[*3] Web 検索エンジンとして，日本語版の Google 検索を用いた.
[*4] 作成手順は文献 [4] に示す.
[*5] 2013 年 1 月 8 日に収集を行った.

表 1:「文化差あり」に分類された語句（83 語）

バレンタインデー	羽根突き	アニメ	醤油	地震	忍者
インターネット	公衆便所	おたく	宗教	豆腐	教員
ファッション	コンビニ	マナー	赤飯	剣道	俳句
おせち料理	中華まん	ビール	貨幣	元日	舞妓
カレーライス	カラオケ	七五三	煎餅	柏餅	漫画
自動販売機	タクシー	わさび	神社	納豆	漫才
クリスマス	ニート	親子丼	神主	祝日	味噌
おにぎり	ちくわ	軍隊	神道	政府	萌え
パチンコ	こたつ	昆布	牛丼	老人	禅
すき焼き	ちまき	空気	相撲	饅頭	茶
就職活動	天ぷら	団子	大人	妖怪	能
ひな祭り	天津飯	干支	浴衣	節分	祭
ラーメン	年賀状	漢字	寿司	首相	侍
アイドル	和菓子	七夕	刺身	落語	

※語句の分類は，日本に 3 年以上滞在している中国人留学生 5 名が行った．各自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．

表 2:「文化差なし」に分類された語句（31 語）

ようかん	化粧品	ケーキ	映画	仕事	母親	鬼
ステーキ	外国人	ゲーム	観光	子供	虹	
固定電話	公務員	花火	教授	戦争	銃	
新婚旅行	時刻表	学歴	発酵	新聞	神	
携帯電話	障害者	友情	父親	大豆	夫	

※語句の分類は，日本に 3 年以上滞在している中国人留学生 5 名が行った．各自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．

提案手法により，114 語中 111 語[*1] に関して，731 件の文化差データを収集した．1 語句あたり平均 6.4 件の文化差データとなった．表 3 に，収集された文化差データの例を示す．表 3 の例では,「日本では」「中国では」「インターネット」が含まれた文が抽出されている．

収集された文化差データの適切さ（文化差データの提示により，文化差の有無を適切に判定可能か）を評価するために，本学のシステム工学部および大学院システム工学研究科の学生 10 名に評価を依頼した．各語句に対する文化差データを見て，各語句の文化差の有無を次の 5 種類で評価してもらった．

- 文化差があると思う
- 少し文化差があると思う
- 文化差はないと思う
- 全て文化差とは無関係の内容だと思う
- わからない

表 4 に，正解データの「文化差あり」「文化差なし」の件数，収集された文化差データの件数，文化差データの正解率を示す．なお，上述した 5 種類への評価結果については,「文化差があると思う」「少し文化差があると思う」と判定された文化差データを「文化差あり」，それ以外を「文化差なし」とした．

「正解率（個別）」は，収集された個別の文化差データによる平均正解率である．正解率は,「文化差あり」「文化差なし」で，それぞれ 0.52，0.45 であり，収集された文化差データの約半分に，適切に文化差判定可能な内容が含まれていることがわかった．

「正解率（各語句）」は，各語句の文化差データ（複数）に対する評価結果を統合し，正解かどうか判定した場合の平均正解率である．各語句の個別の文化差データの評価結果の中に，一つ以上「文化差あり」と判定された文化差データがある場合に,「文化差あり」と判定した．正解率は,「文化差あり」「文化差なし」で，それぞれ 0.81，0.23 であった.「文化差あり」の文化差データの多くを，適切に「文化差あり」と評価できることがわかった．なお,「文化差なし」の語句については，正解率が 0.23 であり，適切な評価ができていない．ただし，本研究では「文化差なし」を誤って「文化差あり」と判定する場合には，大きな問題がないと考えている．

[*1]「友情」「浴衣」「落語」に関する文化差データは 0 件であった．

表 3: 収集された文化差データ（「インターネット」）

日本では ブロードバンドが普及し快適な **インターネット** 接続環境が整備されました。**中国では** 企業にも ADSL が普及しておりますが、固定 IP アドレスは日本に比べ高価です。**中国では** 固定 IP アドレスが高価な地域もありますのでコスト削減になります。**中国のインターネット** は、キャリアの事情によりある特定の経路のみ遮断されることがあります。

表 4: 文化差データの収集結果

	正解データ（件）	収集された文化差データ（件）	正解率（個別）	正 解 率（各語句）
文化差あり	83	547	0.52	0.81
文化差なし	31	184	0.45	0.23
合計	114	731	0.50	0.65

※表中の正解率は，評価者 10 名の平均値である．

本稿で提案した文化差データの収集手法は，個別の文化差データでは，約半分に文化差判定可能な内容が含まれる．特に，複数の文化差データ（今回は最大 20 件）をまとめた場合には，約 8 割の語句は適切に文化差判定が可能であり，本手法により，適切な文化差データを十分に収集可能であると考えられる．

4 おわりに

本稿では，Web 上に存在する「文化差の理解を促す事例」を文化差データとして収集する手法を提案した．評価の結果，複数の文化差データをまとめた場合には，約 8 割の語句は適切に文化差を判定可能であり，本手法により，適切な文化差データを十分に収集できると考えられる．今後は，収集精度の向上について検討し，文化差可視化の Web サービスとして提供を行う．

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「検索エンジンと機械翻訳を用いた多言語用語間における文化差検出サービス」および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] 吉野 孝，宮部真衣：文化差理解支援における可視化効果の検証，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，Vol.2013-GN-86，pp.1-8（2013）.
[3] 久米昭元，長谷川典子：ケースで学ぶ異文化コミュニケーション―誤解・失敗・すれ違い，有斐閣（2007）.
[4] 吉野　孝，宮部真衣：日本語版 Wikipedia における執筆者の意図に着目した日中間の文化差検出手法の検討，電子情報通信学会技術報告，言語理解とコミュニケーション（第 3 回集合知シンポジウム），Vol.111，No.427，NLC2012-58，pp. 13-18（2012）.