1C3-2in

日中質問回答サイトの比較対照分析および文化間差異発見支援

Comparative Analysis of Japanese/Chinese Question Answer Sites and Semi-Automatic Detection of Cross-Cultural Gaps

聶添[*1] 新井翔太[*1] 宇津呂武仁[*2] 河田容英[*3]
Tian Nie Shota Arai Takehito Utsuro Yasuhide Kawada

[*1]筑波大学大学院システム情報工学研究科 Grad. Sch. Sys. & Inf. Eng., Univ. of Tsukuba
[*2]筑波大学システム情報系 Fclty. Eng, Inf. & Sys., Univ. of Tsukuba
[*3](株) ログワークス Logworks Co., Ltd.

This paper focuses on topics such as rabies and raw fish with parasite inside, where people in Japan and China may have different impressions and concerns, mainly because Japan and China have quite different cultural backgrounds and the degree of social advancement especially regarding those issues. Among the types of text data which can be collected from the Web and other information sources, this paper concentrates on questions and answers collected from question-and-answer sites, and claims that those questions and answers are quite informative knowledge source for the task of detecting social and cultural differences between Japan and China. Then, we show how to analyze Japanese and Chinese questions and answers and then show how to detect social and cultural differences between Japan and China.

1. はじめに

日本と中国の間には，過去の歴史に起因する理由などもあり，様々な問題において相反する意見や見解が多い．しかし，近年では，二国間の経済や文化の交流も促進され，お互いの国についての関心が高まりつつあるので，日中間の文化間差異を発見することが大切だと考えられる．ここで，これまで，その国独特の文化の実態は，その国に直接行ってみなければ知ることが難しいのが実情であった．しかし，多くの人にとっては，時間的制約や経済的制約があるため，インターネットやテレビを通して間接的にその国の文化を体験することしかできなかった．また，従来より，文化間の差異の発見に関する研究を実施するためには，一次資料の分析やヒアリング調査などの手法を適用する必要があったが，このような状況においては，それらの研究を実施するためには莫大なコストが必要となっていた．

ところが，近年，情報技術の進展に伴い，インターネットを利用することによって，各国の文化の実態の一端を把握することが可能になりつつある．そこで，本論文では，インターネット上においても，特に，各国の質問回答サイトを情報源とすることによって，直接その国を訪問することなく各国の文化の実態についての情報を収集し，二国間の文化間差異の発見を支援する方式の確立を目的とする．具体的には、本論文においては，日本語と中国語の質問回答サイトから，ある同一の問題についての日本語と中国語の質問と回答を収集し，回答中の記述を対象として，日本と中国の間の文化的対照性を分析する方法を提案する．ここで，我々は，[中崎 10] において，日英ブログサイトにおける文化間差異の発見を支援する方式を提案したが，本論文では，日中質問回答サイトにおける文化間差異発見支援のタスクを対象としてこの方式を適用する．本論文において提案する「日中質問回答サイトの比較対照分析による文化間差異発見支援」の枠組みを図 1 に示す．

まず，本論文では，日本語質問回答サイトとして Yahoo!知恵袋[*1] を，また，中国語質問回答サイトとして，Baidu(百度)知道 [*2]，および，搜搜問問 [*3] を，それぞれ利用する．そして，特定の話題を表すクエリを用いて，日中各言語の質問回答サイトから質問・回答組を収集し分析を行う．分析の第 1 ステップにおいては，日中各言語において 100∼200 組程度の質問・回答組を対象として，質問・回答組の内容を分析し，類似する内容のまとめ上げを行い，異なる内容の数を集計する．次に，日中の間でその内容の対応関係をつけ,「日本側でのみ観測した内容」,「中国側でのみ観測した内容」,「日中両側で観測した内容」に分類して集計する．次に，分析の第 2 ステップにおいては,「日本側でのみ観測した内容」，および,「中国側でのみ観測した内容」を対象として，当該内容が観測されなかったと判定された言語側で追加の検索を行い，本当にその当該内容が検索されないかどうかの検証を行う．

例えば，図 1 に示すように,「犬に噛まれた」という話題を対象とした場合，中国語の質問回答サイトにおいて最も多い回答としては,「ワクチンを接種したほうがよい」といった緊張感のある回答の件数が圧倒的になるのに対して，日本語の質問回答サイトにおいて最も多い回答としては,「消毒をすれば大丈夫」といった緊張感の少ない回答の件数が圧倒的となる．この背景には，中国は狂犬病発症率が世界第二位であるのに対して，日本では 1,956 年以来狂犬病の発生が記録されていない，といった社会的文化的差異の影響が大きく関わっているということができる．

2. 日中質問回答サイトの質問・回答事例

日本側の質問回答サイトの質問・回答事例としては，Yahoo! 知恵袋から提供されている 2004 年 4 月 1 日 ∼2009 年 4 月 7 日の 5 年間の質問・回答事例のデータ (質問: 16,257,413 件, 回答: 50,053,894 件) を分析対象とした．質問には，カテゴリ情報が付与されており，最下位層の分類として 453 種のカテ

連絡先: 聶 添 ，筑波大学大学院システム情報工学研究科，
〒 305-8573 茨城県つくば市天王台 1-1-1, 029-853-5427

*1 http://chiebukuro.yahoo.co.jp/
*2 http://zhidao.baidu.com/
*3 http://wenwen.soso.com/

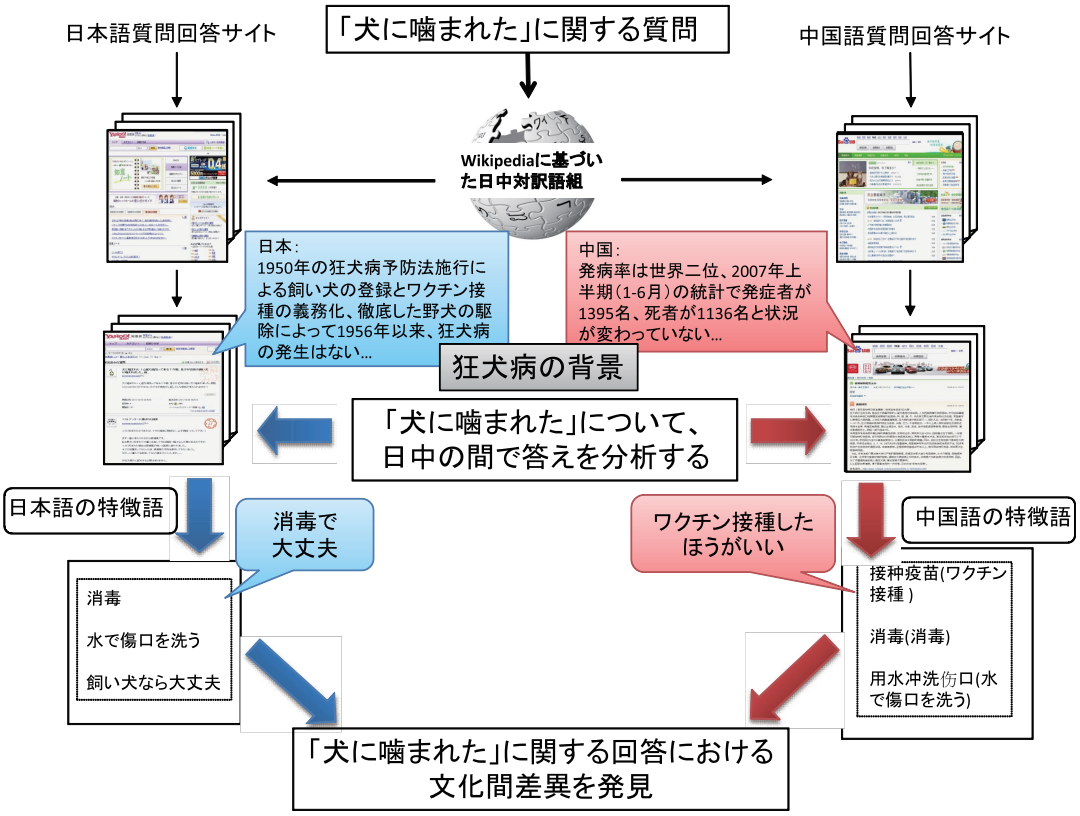

図 1: 日中質問回答サイトの比較対照分析による文化間差異発見支援

表 2: 日中間における質問回答組の内容の比較: 第 1 ステップ / 第 2 ステップ

話題		日本側でのみ観測した内容の数	中国側でのみ観測した内容の数	日中両側で観測した内容の数
第 1 ステップのみ	犬に噛まれる	10 / —	6 / —	9 / —
	刺身と寄生虫	10 / —	9 / —	1 / —
第 1/第 2 ステップ	喫煙	10 / 6	13 / 9	5 / 13
	原発	8 / 7	3 / 2	4 / 6

表 1: 各話題において用いたクエリ・クエリの出現数・分析対象質問・回答組数

話題	日本側			中国側 (Baidu / 搜搜)		
	クエリ	クエリの出現数	分析対象質問・回答組数	クエリ	クエリの出現数	分析対象質問・回答組数
犬に噛まれる	犬に噛まれた	136	124	被狗咬了	743/499	124/135
刺身と寄生虫	刺身 and 寄生虫	89	87	刺身 and·寄生虫	307/調査中	76/—
原発	原発	1,412	131	原子能发电	750/調査中	93/—
喫煙	喫煙	18,534	149	吸烟	741/344	141/149

ゴリが存在している．453 種のカテゴリは，それぞれ親カテゴリ，さらにその親カテゴリを持つ三層構造になっており，各カテゴリに数万 ~ 数十万の質問が含まれている．

一方，中国側の質問回答サイトの質問・回答事例は，2012 年 12 月 ~2013 年 2 月の期間に，Baidu(百度) 知道，および，搜搜問問のサイトから収集した．2012 年 12 月の時点で，Baidu(百度) 知道に掲載されていた解決済質問数は 215,755,535 件，搜搜問問に掲載されていた解決済質問数は，207,374,517 件であった．Baidu(百度) 知道，および，搜搜問問のサイトのいずれも，質問には 3~4 階層のカテゴリ情報が付与されており，最上位層のカテゴリ数は 14 種類であった．

3. 分析対象の質問事例の収集

本論文では，特に，日本と中国の間で文化的な差異が大きいことが期待できる話題として，以下の 4 つの話題を選定した．以下では，各話題およびその選定理由を示す．

- 「犬に噛まれる」··· 1. 節で述べた通り．
- 「刺身と寄生虫」··· 中国では生魚の衛生管理面での信頼性が日本よりも低いと考えられている．
- 「原発」··· 日中両国においては，原発への関心の度合いに差異がある (ただし，日本側において分析対象とした質問・回答事例は 2009 年時点までのものである)．
- 「喫煙」··· 日本では 20 歳未満は喫煙禁止であり，公共の場所での喫煙は禁止されているが，中国では喫煙できる年齢に関する法律はなく，公共の場所での喫煙が禁止されたのも最近のことである．

表 3: 話題「犬に噛まれた」において観測された内容および質問・回答組数 (第 1 ステップのみ)

	話題	日本側の質問・回答組数	中国側の質問・回答組数 (Baidu/搜搜)
日本側でのみ観測	**飼い主の責任の所在に関する相談 (9 件)** **近所との人間関係についての相談 (8 件)** **飼い犬に噛まれたことがある回答者への相談 (8 件)** **ペットと飼い主との関係についての相談 (7 件)** **破傷風の往診を勧める (7 組) / 犬に噛まれても大丈夫 (6 組)** **犬に噛まれた傷跡についての相談 (6 件)** 犬のワクチン接種を勧める / 狂犬病の保険についての相談 飼い犬が人を噛まない理由についての質問	計 57	—
中国側でのみ観測	**噛んだ犬の様子を観察し狂犬病の犬かどうか判断することを勧める (18 件)** 出血がなければ大丈夫 / 潜伏期間が過ぎれば大丈夫 狂犬病のワクチンの値段に関する質問 犬に噛まれたことについて悩んでいる精神疾患についての相談 ワクチンを接種した犬に噛まれた場合は狂犬病になるとは限らない	—	計 26 / 計 22
日中両側で観測	**狂犬病ワクチンの接種を勧める**	1	**68 / 72**
	狂犬病の基礎知識を説明し，必要があればワクチン接種を勧める	1	**15 / 33**
	狂犬病の基礎知識を説明	2	6 / 1
	噛まれたことを心配している質問に対し対策を回答しているが，回答者に狂犬病の知識はない.	**28**	4 / 3
	犬に噛まれたことに関する質問および回答において法律関係の事項を含む	**20**	1 / 0
	ペットが犬に噛まれたのでアドバイスを求めている (狂犬病への言及なし)	6	3 / 1
	犬恐怖症対策に関する相談	5	1 / 0
	犬に噛まれた時の感触についての相談	3	0 / 1
	狂犬病の往診を勧める	1	0 / 2
合計		124	124 / 135

そして，各話題について，表 1 に示す日中両言語のクエリを用いて，各クエリを文字列として含む質問・相談組を収集した．特に，Baidu(百度) 知道，および，搜搜問問においては，収集可能な質問・回答組の上限が，それぞれ，750 組，および，500 組であったので，この上限の範囲内で収集可能な質問・回答組を収集した．収集された質問・回答組数を，表 1 中の「クエリの出現数」欄に示す．次に，各クエリを文字列として含む質問・回答組を人手で分析し，分析対象とする話題に関連する質問・回答組を選別し，最終的に表 1 中の「分析対象質問・回答組数」欄に示す個数の質問・回答組を分析した．

4. 日中文化間差異の分析

4.1 分析手順

第 1 ステップ まず，前節で収集された日中両言語の質問・回答組の内容を人手で分析し，類似の内容の質問・回答組をまとめることにより，内容の種類数を集計する．その後，それらの内容を日中間で対応付けることにより，

(1) 日本側でのみ観測した内容,

(2) 中国側でのみ観測した内容,

(3) 日中両側で観測した内容,

に分類する．分析対象とした 4 つの話題について，上記の (1)∼(3) の内容の数を表 2 に示す．

第 2 ステップ 次に，話題「喫煙」，および,「原発」について,「(1) 日本側でのみ観測した内容」が中国側の Baidu(百度) 知道，および，搜搜問問において観測できないかどうか，収集した全質問・回答組 (「原発」については，750 組 (Baidu(百度) 知道),「喫煙」については，741 組 (Baidu(百度) 知道)，および，344 組 (搜搜問問)) を対象として調査を行う．同様に,「(2) 中国側でのみ観測した内容」が,日本側の Yahoo!知恵袋において観測できないかどうか,収集した全質問・回答組 (「原発」については，1,412 組,「喫煙」については，18,534 組) を対象として調査を行う．分析対象とした 2 つの話題について，この第 2 ステップの後の (1)∼(3) の内容の数を同様に表 2 に示す．

4.2 分析結果

表 2 に示すように，4 つの話題のいずれにおいても,「(1) 日本側でのみ観測した内容」，および,「(2) 中国側でのみ観測した内容」が一定数観測されていることが分かる．

次に，話題「犬に噛まれた」について，上述の (1)∼(3) の内容の詳細，および，該当する質問・回答組数を表 3 に示す．表中では，同一の内容について 6 件以上の質問・回答組数が観測できた場合に，その内容を**太字**で示す．この結果から分かるように，日中両側で観測された内容「狂犬病ワクチンの接種を勧める」，および,「狂犬病の基礎知識を説明し，必要があればワクチン接種を勧める」に該当する件数が，中国側では圧倒的多数を占めるのに対して，日本側では 1 件のみとなっており，日中間の関心の度合いの違いが，狂犬病の危険度の違いを如実に反映した結果となった．同様に，中国側でのみ観測され

表 4: 話題「刺身と寄生虫」において観測された内容 (第 1 ステップのみ，抜粋)

日本側でのみ観測された内容	中国側でのみ観測された内容	日中両側で観測された内容
特定の魚の刺身が心配という相談に対して，問題ないという回答 (42 件) **特定の魚の刺身が心配という相談に対して，魚種によっては寄生虫が危険という回答 (14 件)** 生の魚は心配という相談に対して，食べない方がよいという回答 (5 件) その他 26 件の大半は「刺身は安全」という回答	**魚の刺身が心配という相談に対して，寄生虫が危険なので食べない方がよいという回答 (50 件)** **魚の刺身が心配という相談に対して，海水魚なら問題ないという回答 (11 件)** その他 15 件の大半は「生魚を不安視」する回答	川魚の刺身についての相談に対し，食べない方がよいという回答 (日本 4 件，中国 2 件)

表 5: 話題「喫煙」において観測された内容 (第 1 ステップ・第 2 ステップ)

(a) 日本側または中国側でのみ観測された内容

日本側でのみ観測	中国側でのみ観測
喫煙者が禁煙の風潮に対するクレームを相談 (27 件) **喫煙者に対する印象についての質問と回答 (悪い印象であるという回答 (25 件)**，中立的な印象であるという回答) 喫煙者と非喫煙者が共存する仕組みについて相談と回答 喫煙を始めた時期についての質問と回答 国会議員の喫煙に関する発言についての議論	**喫煙のよい点についての質問と回答 (喫煙者は魅力的，喫煙すると集中できる，国の税収増) (29 件)** **喫煙の仕方についての質問と回答 (13 件)** 中国の喫煙率についての質問と回答 (2009 年当時で約 25%) 喫煙した方が大人びて見えるかという質問と回答 (関係ないと回答) その他，3 種類の内容

(b) 日中両側で観測された内容

喫煙の害についての質問と回答 (日本 18 件，中国 215 件) / 同僚の喫煙による被害の回避策について相談 (日本 37 件) その他，10 種類の内容

た内容としては,「噛んだ犬の様子を観察し狂犬病の犬かどうか判断することを勧める」があり，しかもその件数は圧倒的多数となった．一方，日本側では，狂犬病以外の「法律関係の事項」,「飼い主の責任の所在」,「近所との人間関係」,「ペットと飼い主の関係」，といった観点での質問・回答組が多数を占めることが分かる．

同様に，話題「刺身と寄生虫」について，上述の (1)〜(3) の内容のうち主なものの抜粋を表 4 に示す．この結果から分かるように，日本では，通常家庭や外食店で食べられる刺身については寄生虫の心配はないが，魚種によっては寄生虫が危険な種類もあり，専門的な知識が必要である，といった回答が多数派を占め，社会全体が十分に管理されているという印象を持つが，中国では,「生の魚は寄生虫が危険である」という回答が多数派を占めるという結果であった．

また，話題「喫煙」についても，同様の抜粋を表 5 に示す．

5. 関連研究

本論文の先行研究として，我々は，[中崎 10] において，特定の話題について，日本語ブログ記事，および，英語ブログ記事を収集し，関心事項や賛否に関する文化間差異発見過程を支援する方式を提案した．この方式の成果として,「捕鯨」や「臓器移植」など，日本と欧米圏との間で社会制度上の違いや食文化の差異が大きい話題について，ブログ空間における関心の違いを容易に観測することができた．また，[Yoshioka08] は，複数の国の代表的なメディアが発信するニュースを情報源として，同一事象に対する各国のニュースの伝え方の差異分析方式を提案している．その他，[胡 13] においては，日中の時系列ニュースを対象として，時系列トピックモデルを適用し，日中単言語のトピックの間の言語間対応をとることにより，同一の話題に関するニュース記事の集合を持つ日中各言語のトピックを同定する方式を提案している．

6. おわりに

本論文では，日本語と中国語の質問回答サイトから，ある同一の問題についての日本語と中国語の質問と回答を収集し，回答中の記述を対象として，日本と中国の間の文化的対照性を分析する方法を提案した．今後は，特定の話題について収集した日中両言語の質問・回答組の間で，専門的内容を表す用語の対訳関係を利用することにより，日中間の文化間差異の有無を自動判定する方式について研究を行う．

謝辞

本研究においては，ヤフー株式会社より提供して頂いた Yahoo!知恵袋のデータを利用させて頂いた．関係各位に感謝の意を表する．

参考文献

[胡 13] 胡碩, 高橋佑介, 鄭立儀, 宇津呂武仁, 吉岡真治, 神門典子：日中時系列ニュースにおけるバースト・トピックの推定と二言語間対応付け, 言語処理学会第 19 回年次大会論文集, pp. 204–207 (2013).

[中崎 10] 中崎寛之, 川場真理子, 横本大輔, 宇津呂武仁, 福原知宏：多言語 Wikipedia エントリを知識源とする特定トピックの日英ブログサイト検索と日英対照ブログ分析, 人工知能学会論文誌, Vol. 25, No. 5, pp. 613–622 (2010).

[Yoshioka08] Yoshioka, M.: IR Interface for Contrasting Multiple News Sites, *Prof. 4th AIRS*, pp. 516–521 (2008).