E-14

文化差判定コーパス構築におけるクラウドソーシング利用の検討

Discussion of Building Cultural Differences Judgment Corpus Using Crowdsourcing

吉野　孝 †　　宮部真衣 ‡
Takashi Yoshino　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]. 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，文化差検出に関する検討を行ってきた．語句に関する文化差の有無に関しては，Wikipedia の項目数に基づく手法，記事内の国名・言語名数に基づく手法，記事における執筆者の意図に基づく手法を統合することにより，比較的高い精度での判定を可能にした [2].

しかし，文献 [2] で用いた評価データは，中国人留学生の協力をもとに作成した評価データであり，作成におけるコストが高く，評価語句数も 114 語と少ない．文化差検出手法を評価するためには，ある程度の規模の評価データ（文化差判定のためのコーパス）が必要となる．

そこで我々は，クラウドソーシング（Crowdsourcing）[3] を利用した文化差判定コーパスの構築を検討した．本稿では，クラウドソーシングを利用した文化差判定コーパス構築のために行った実験とその結果および考察を示し，クラウドソーシングの利用可能性について議論する．

2. 関連研究

クラウドソーシングは，単純な作業であるが，計算機を用いて自動化することが困難な作業，あるいは計算機に比べて人間が得意な作業を，インターネットを介して不特定多数の人に業務委託する方法である [3]．クラウドソーシングは多人数による処理のため，大量のデータを短時間で処理することができたり，安価で作業が可能なため，全体的なコストを低減させたりすることができる．しかし，不特定多数の人が作業者となるため，作業品質が低いという問題がある．

クラウドソーシングを用いた研究は多数行われている．例えば，Jha らは，Amazon Mechanical Turk を利用し，ブログなどの非定型で省略や誤りの多い文から，前置詞句の正確な起点・着点等を抽出し，コーパスを作成する実験を行った [4]．実験の結果，高精度なコーパスの構築が可能であることを示した．

芦川らは，新語や未知語の単語の「読み」「アクセント」などの付与に，独自のクラウドソーシングを利用した [5]．芦川らは，同一タスクを用いて，Amazon Mechanical Turk と独自のクラウドソーシングとの精度の比較を行い，芦川らのシステムが，Amazon Mechanical Turk より高精度な結果を得られることを示した．

文化差判定コーパスは，著者らの文化差検出手法を評価するためのデータであり，既存のコーパスは存在しない．本研究では，コストの掛かる文化差判定コーパスの構築にクラウドソーシングを利用することを検討している．

3. 文化および文化差の定義

本章では，本稿で対象とする「文化差」の定義について述べる．

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 [6]．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 [6]．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,「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」のある内容は，一方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，別の文化圏にも伝わっている場合も，基本的には，もとの文化圏の内容を指しているものである．例えば，日本の地名のいくつかは，海外にも伝わっているが，その地名の持つ背景（歴史的あるいは文化的）などは正確には理解されない.「第 2 種の文化差」のある内容は，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．例えば,「醤油」は日本と中国のどちらにも存在するが，日本の「醤油」と中国の「醤油」は異なる[*1]．

4. 文化差判定コーパス構築におけるクラウドソーシングの利用

本章では，クラウドソーシングを用いた文化差判定コーパス構築手法について述べる．

我々はこれまでに，文化差検出手法の精度評価のために，文化差判定コーパスの構築を行った [2]．構築したコーパスには，日本と中国の文化差のある語句が含まれる．文化差判定コーパスは，以下の手順により構築した．

作業 1 語句の収集

存在するあらゆる語句に対して，文化差の有無が判定されるのが理想的であるが，語句の数は膨大であり，全語句の判定は困難である．文化差判定コーパスは，文化差検出手法の評価に用いるものであり，評価のための語句としては，文化差のある語句が含まれる必要がある．そこで，一方の文化圏（今回は日本）の作業者により，文化差があると考えられる語句を収集する．

† 和歌山大学システム工学部，Faculty of Systems Engineering, Wakayama University
‡ 東京大学知の構造化センター，Center for Knowledge Structuring, The University of Tokyo

[*1] 中国の醤油は，味やにおいが日本の醤油よりも強い．中華料理での使われ方は，香りや味よりも，色づけに重点を置いている．日本の醤油は，中国から日本に伝わり，日本で独自に改良されたと言われており，製造方法も異なる．また，世界各国に輸出あるいは現地生産されており，広く受け入れられている．

表 1: 正解データの分布

文化差の種類	表記	件数
文化差なし	Type 0（中国にもあって，日本と同じ）	31
第 1 種の文化差	Type 1（中国にはない）	42
第 2 種の文化差	Type 2（中国と日本のものは違う）	41
合計		114

表 2: 日本語と中国語の文化差判定済みのデータの例

文化差の種類	例
Type 0（中国にもあって，日本と同じ）	携帯電話，化粧品，公務員，ケーキ
Type 1（中国にはない）	牛丼，おたく，おせち料理，赤飯，すき焼き
Type 2（中国と日本のものは違う）	昆布，干支，アイドル，タクシー，地震

図 1: 文化差判定のための Worker 用の Web 画面の例

作業 2 文化差の判定

作業 1 で収集した語句に対して，文化差の判定を行う必要がある．文化差の有無については，各国（今回は日本および中国）の文化に関する知識を持った人間でなければ，正しく判定できないと考えられる．そこで，各国の知識を持った作業者（今回は中国人留学生）に，文化差の有無を判定してもらう．また，作業者の判定結果をもとに，文化差の種類（第 1 種の文化差，第 2 種の文化差，文化差なし）を決定する．

作業 1 は，一方の文化圏のみを理解していれば作業可能であるため，作業者の能力を問わない比較的容易な作業である．一方，作業 2 は，各国の文化に関する知識が必要であり，作業者が限定される．特殊な能力が要求されるため，まず，作業 2 を行うことができる作業者を集めることが容易ではない．また，判定の正確さは知識量に依存し，判定結果に個人差が発生するため，複数名による判定を行う必要がある．そのため，文化差判定コーパスの構築においては，特に作業 2 において高いコストがかかる．

そこで我々は，これらの作業 2 に関する問題点を，クラウドソーシングにより解決できる可能性があると考えた．2. 章で述べたように，クラウドソーシングは，インターネットを介して不特定多数の人に業務委託でき，またコストの低下が期待できる．つまり，上述した作業者の確保や，コストの問題については解決可能であると考えられる．しかし，クラウドソーシングでは，作業品質が低くなる可能性がある．

そこで本稿では，クラウドソーシングによる作業 2 の作業精度を検証し，クラウドソーシングによる文化差判定コーパス構築作業の代替可能性について，作業品質の面から検討する．今回は，下記の手順により，クラウドソーシングによる文化差判定結果を検証する．

Step 1 日本語の語句と中国語の語句を，クラウドソーシングを利用して，評価する．

Step 2 評価結果をもとに,「第 1 種の文化差」「第 2 種の文化差」「文化差なし」を決定する．

5. 検証実験

本章では，クラウドソーシングによる文化差判定実験および実験結果について述べる．

5.1 実験方法

本稿では，クラウドソーシングサービスとして，CrowdFlower[*1] を用いて，Amazon Mechanical Turk[*2] を利用した[*3]．提案した文化差判定コーパス構築手法の適用可能性を調べるために，下記の実験を行った．

(1) 日本語と中国語の文化差判定済みの語句（114 件）の判定を，クラウドソーシングサービスに依頼する[*4]．表 1 に文化差判定済みのデータの分布を示す．それぞれ，30 件〜40 件程度のデータがある．表 2 に，各語句の一部を示す．なお，文化差の種類の表記は，Type 0，Type 1，Type 2 とし，それぞれ,「文化差がない」,「第 1 種の文化差」,「第 2 種の文化差」を表す．

(2) 判定基準は下記の 4 通りである．

- 中国にはない（第 1 種の文化差）
- 中国にもあって，日本と同じ（文化差なし）
- 中国と日本のものは違う（第 2 種の文化差）
- わからない

(3) 図 1 に，Amazon Mechanical Turk 上の評価画面を示す．英語と日本語を併記し，さらに中国語の語句は中国語（簡体字）で提示している．判定基準「中国と日本のものは違う」を選択した場合には，違う理由を記載するようにしている．

(4) CrowdFlower では，評価データの中に Gold Units という正解データを一部含めることによって，Worker の質を保つ仕組みがある．今回は，12 件の Gold Units を設定した．

5.2 実験結果

当初，1 語句の判定に対して，5 セント支払うように設定していたが，作業者（Worker）が約 3 日間経過しても十分に集

[*1] http://crowdflower.com/

[*2] https://www.mturk.com/

[*3] Amazon Mechanical Turk への直接の仕事の依頼は，アメリカ合衆国在住者に限られているため，CrowdFlower を介して利用した．

[*4] 文化差判定済みのデータは，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および日本に 3 年以上住んでいる本学の中国人留学生 5 名に依頼し，分類を行った [2].

表 3: Worker の正解率

Worker ID	正解数	回答数	正解率
A	41	108	38%
B	40	104	38%
C	47	108	44%
D	41	114	36%
E	41	108	38%

・評価データの総数は 114 個であるが，Worker によっては評価漏れがあったため回答数は異なる．

表 4: 正解データと Worker の判定結果との相関係数

Worker ID	相関係数
A	0.106
B	0.073
C	0.013
D	0.009
E	0.106

まらなかったため[*1]，1 語句の判定に対して 25 セント支払うように設定したところ，約 3 日間で 5 名の Worker が評価を完了した[*2]．5 名の Worker の Gold Units の正解率は，90 %以上である．CrowdFlower のアクセスログによると，全ての Worker は，中国から作業を行っている．

Worker の正解率について

表 3 に，各 Worker の正解率を示す．評価データの数は 114 件だったが，Worker が分割して作業した場合に，複数回 Gold Units が表示されるため，回答数が 114 でない場合がある．

表 3 から，正解率は 36%から 44%であり，作成されたコーパスを単純に利用するには十分な精度が得られていない．また，判定基準で「中国と日本のものは違う」を選択した場合には，違う理由を記載するようにしているが，あまり記載されていなかった．

正解データと各 Worker の判定結果との相関

表 4 に，正解データと Worker の判定結果との相関係数を示す．相関係数は 0.013 から 0.106 であり，相関はほとんどないことがわかった．

Worker 間の相関

表 5 に，Worker 間の相関係数を示す．WorkerA，WorkerB，WorkerD の相関係数は 0.762〜0.856 であり，かなり高いことが分かった．

Worker の判定結果の分布

表 6 に，Worker の判定結果の分布を示す．相関の高かった WorkerA，WorkerB，WorkerD の判定結果の分布が近い．また，WorkerC と WorkerE は，Type 0 の判定が特に多い．

Worker の適合率，再現率，F 値

表 7，表 8，表 9 に，それぞれ，Worker の判定結果の適合率，再現率，F 値を示す．

表 7 をみると，WorkerC および WorkerE の Type 1 の適合率が高い．表 6 から，WorkerC および WorkerE は，Type 1 と判定した語句は，それぞれ 17 個，8 個と，その他の Worker に

[*1] 10 名の Worker が作業を開始したが，7 名が 20 語句までで評価を止めていた．60 語句以上を評価した Worker は 0 名だった．

[*2] 24 名の Worker が作業を開始し，16 名が 20 語句までで評価を止めていた．100 語句以上を評価した Worker は 5 名だった．

表 5: Worker 間の相関係数

	A	B	C	D	E
A	1.000	0.856	0.011	0.762	0.261
B	-	1.000	0.076	0.788	0.186
C	-	-	1.000	0.060	0.172
D	-	-	-	1.000	0.331
E	-	-	-		1.000

表 6: Worker の判定結果の分布

	Type 0（中国にもあって，日本と同じ）	Type 1（中国にはない）	Type 2（中国と日本のものは違う）	合計
A	42	48	18	108
B	40	46	18	104
C	83	17	8	108
D	43	46	25	114
E	92	8	8	108
平均	60	33	15	

・評価データの総数は 114 個であるが，Worker によっては評価漏れがあったため回答数は異なる．

比べて少なく，これらの多くが正解となっているため，Type 1 の適合率が高いことが分かった．

表 8 をみると，WorkerC および WorkerE の Type 0 の再現率が高い．表 6 から，WorkerC および WorkerE は，多くの語句を Type 0 と判定しているため，正解も多く含まれており，Type 0 の再現率が高いことが分かった．

表 9 を見ると，F 値の平均は特に大きな差が見られなかった．また，WorkerC の Type 1 および WorkerE の Type 2 の F 値が高い．

Worker の間違いの傾向

表 10 に，Worker が間違って判定した文化差を示す．「Type 1 および Type 2 を Type 0 に間違って判定」が，平均 62 %あり，間違った判定の多くは Type 0 と判定されていることが分かった．「Type 0」は，「中国にもあって，日本と同じ」であり，日本にもあることが分かれば，Type 0 と判定しやすい可能性がある．

表 11 に，各分類における Worker の誤判定率を示す．「Type 2 の判定を間違えた」が，平均 49 %であり，多くの Type 2 の判定を間違えていることが分かった．「Type 2」は，「中国と日本のものは違う」であり，単に日本にあることが分かっても，中国と日本の文化との違いが分からない場合には，正しく判定することが難しいと考えられる．

5.3 多言語を取り扱う際のクラウドソーシング利用の限界

今回，文化差判定コーパス構築のために，クラウドソーシングを利用した実験を行った．今回の実験結果から，多言語を取り扱う際のクラウドソーシング利用の限界について議論する．

多言語を取り扱う際のクラウドソーシングとしては，「人力翻訳プラットフォーム」Gengo[*3]がある．Gengo は，資格審査を経た数千人の翻訳者が登録されており，安価で高速な翻訳サービスを提供している．しかし，文化差判定コーパス構築に関しては，上記の翻訳サービスは利用出来ない．

今回の実験結果から，クラウドソーシング上（少なくとも Amazon Mechanical Turk 上）には，複数言語を理解する Worker（少なくとも日本語と中国語を両方とも理解する Worker）は，

[*3] http://gengo.com/，2012 年 7 月に myGengo から Gengo に名称を変更した．

表 7: Worker の判定結果の適合率

	Type 0（中国にもあって，日本と同じ）	Type 1（中国にはない）	Type 2（中国と日本のものは違う）
A	0.333	0.396	0.444
B	0.300	0.413	0.500
C	0.325	1.000	0.375
D	0.279	0.413	0.400
E	0.304	0.875	0.750
平均	0.308	0.619	0.494

表 8: Worker の判定結果の再現率

	Type 0（中国にもあって，日本と同じ）	Type 1（中国にはない）	Type 2（中国と日本のものは違う）
A	0.452	0.452	0.341
B	0.387	0.452	0.293
C	0.871	0.405	0.659
D	0.387	0.452	0.293
E	0.903	0.167	0.683
平均	0.600	0.386	0.454

ほとんどいないことが分かった．図 1 に示すとおり，今回は，日本語をベースとした評価画面で評価を行った．著者らが理想としている Worker は，日本語が理解可能な中国人であった．しかし，実際の Worker は，日本語が理解できない中国人の可能性が高いと考えられる．これは,「違う理由」に記載された内容として，中国語や不完全な日本語（機械翻訳により生成された日本語と思われるもの）が含まれていることから推察される．

今回，Worker の質を判断する Gold Units を利用した．各 Worker の Gold Units の正解率は高かったが（90 %以上），文化差判定済みの語句の正解率は低く（表 3），十分に機能していないことが分かった．多くの Worker は，機械翻訳等を利用することで，日本語の質問内容を理解し，評価していると考えられる．

今回，文化差の判定として,「Type 0（中国にもあって，日本と同じ）」「Type 1（中国にはない）」「Type 2（中国と日本のものは違う）」の 3 種類の判定を試みた．表 7 をみると,「Type 1」の平均は他と比べて高い．特に WorkerC および WokerE は高い.「Type 0」および「Type 2」については，中国および日本の両方のことを知らないと判定出来ないが,「Type 1」については，中国のことが分かれば正しく判定できる可能性が高い．クラウドソーシングを十分に有効活用するためには，ある程度の能力が必要とされるタスクを，作業者の言語や文化を問わない単純なタスクへ分解する必要があると考えられる．

6. おわりに

今回，文化差検出システムを評価するための文化差判定コーパスの構築のために，クラウドソーシングを利用した手法について検討した．

本研究の貢献は，文化差判定コーパス構築のためにクラウドソーシングを利用し，多言語を取り扱う際のクラウドソーシング利用の限界について検討を行った点である．

今回，クラウドソーシングの単純な利用では，十分な精度の文化差判定コーパスの構築は難しいことが分かった．しかし，文化差判定コーパスの構築を留学生などに依頼すると大きなコストが掛かる．今後は，作業者の言語や文化を問わないタスクを検討し，クラウドソーシングを利用した文化差判定コーパスの構築を目指す．

表 9: Worker の判定結果の F 値

	Type 0（中国にもあって，日本と同じ）	Type 1（中国にはない）	Type 2（中国と日本のものは違う）
A	0.384	0.422	0.386
B	0.338	0.432	0.369
C	0.474	0.576	0.478
D	0.324	0.432	0.338
E	0.455	0.280	0.715
平均	0.395	0.428	0.457

表 10: Worker が間違って判定した文化差

	Type 1 と Type 2 を Type 0 に間違って判定	Type 0 と Type 2 を Type 1 に間違って判定	Type 0 と Type 1 を Type 2 に間違って判定
A	42%	43%	15%
B	44%	42%	14%
C	92%	0%	8%
D	42%	37%	21%
E	96%	1%	3%
平均	62%	25%	12%

表 11: 各分類における Worker の誤判定率

	Type 0 の判定を間違えた	Type 1 の判定を間違えた	Type 2 の判定を間違えた
A	22%	30%	48%
B	25%	28%	47%
C	2%	39%	59%
D	26%	32%	42%
E	4%	46%	49%
平均	16%	35%	49%

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，諏訪智大：Wikipedia を用いた文化差可視化サービスの開発，情報処理学会，マルチメディア，分散，協調とモバイル（DICOMO2011）シンポジウム (2012).

[3] Jeff Howe: Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business, Crown Business (2008).

[4] Mukund Jha, Jacob Andreas, Kapil Thadani, Sara Rosenthal and Kathleen McKeown: Corpus Creation for New Genres: A Crowdsourced Approach to PP Attachment, Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT'10), pp.13–20 (2010).

[5] 芦川将之，西山　修，下郡信宏：CrowdSourcing を用いた単語への読み付け、及びアクセント付け，電子情報通信学会，人工知能と知識処理，AI2011-36，pp. 11-16 (2012).

[6] 西田ひろ子，異文化間コミュニケーション，創元社 (2000).