情報処理学会論文誌 Vol.59 No.1 69–75 (Jan. 2018)

クラウドソーシング作業者の作業時間の比を用いた多言語用例対訳正確性評価手法

福島 拓[1,a)] 吉野 孝[2,b)]

受付日 2017年4月13日, 採録日 2017年10月3日

概要：正確な多言語間コミュニケーションが求められる場においては，多言語テキストペアである用例対訳が多く用いられている．用例対訳の提供には正確性評価が重要となるが，評価対象の用例対訳は多く，十分な量の正確性評価を集めることが今後さらに難しくなると考えられる．そこで本論文では，クラウドソーシングを用いた用例対訳候補の正確性評価手法を提案する．本手法では，クラウドソーシング上の作業者から作業時間の比を用いて正確に評価する人を優先的に選択することで，より正確性の高い評価手法を提案する．本論文の貢献は，タスク作業時間の比をもとにした作業者の順位付けを行うことにより，従来手法よりも高い正確性を持つ正確性評価手法を提案した点である．

キーワード：クラウドソーシング，用例対訳，多言語間コミュニケーション支援，教師なし分類

Correctness Evaluation Method for Multilingual Parallel Texts Using Ratio of Working Times of Workers via Crowdsourcing

Taku Fukushima[1,a)] Takashi Yoshino[2,b)]

Received: April 13, 2017, Accepted: October 3, 2017

Abstract: Generally, multilingual support systems for applications requiring high accuracy use multilingual texts pairs called “parallel texts.” Accuracy evaluation is important for providing parallel texts. However, because there are huge quantities of parallel texts, it is difficult to thoroughly collect accuracy evaluations. In this paper, we propose a correctness evaluation method for parallel text-candidates using crowdsourcing. This method ranks workers via crowdsourcing based on their ratio of work time. This idea aims to preferentially select workers who evaluate accurately. The major contribution of this study is a correctness evaluation method of parallel text candidates, having a higher accuracy than that of existing methods. Furthermore, the proposed method ranks workers based on their ratio of working time.

Keywords: crowdsourcing, parallel text, multilingual communication support, unsupervised classification

1. はじめに

近年の世界的なグローバル化により多言語間コミュニケーションの機会が増加している．日本国内でも今後，外国人住民のさらなる増加が予想されている [1]．このため，政府内でも多文化共生の推進に関する研究会が開かれており [1]，今後，多文化共生社会になると考えられる．しかし，一般に多言語を十分に習得することは非常に難しく，母語以外の言語によるコミュニケーションは困難なこともあり [2], [3], [4]，日本語を理解できない外国人と日本人との間で正確な情報共有を十分に行うことはできない．

日本語を理解できないことの影響が顕著に現れる分野の1つに医療がある．医療分野では，わずかなコミュニケーション不足で医療ミスが発生する恐れがある．特に，日本語が通じない外国人と日本人の医療従事者間でのやりとりは，意思の疎通を十分に行うことができない．現在，日本

[1] 大阪工業大学情報科学部
Faculty of Information Science and Technology, Osaka Institute of Technology, Hirakata, Osaka 573–0196, Japan
[2] 和歌山大学システム工学部
Faculty of Systems Engineering, Wakayama University, Wakayama 640–8510, Japan
[a)] taku.fukushima@oit.ac.jp
[b)] yoshino@sys.wakayama-u.ac.jp

語を理解できない外国人の支援は医療通訳者が行っているが，医療通訳者は慢性的な人員不足となっている．

そこで，多言語対応の医療支援システムの開発が多く行われている [5], [6]．これらのシステムでは，正確な多言語変換が可能な用例対訳が用いられている．用例対訳とは，用例を多言語に正確に翻訳した多言語コーパスのことを指し，「保険証はお持ちですか？」「はい」「いいえ」などの利用現場で使用される言葉を多言語で提供することができる．この用例対訳を用いて，利用者が適切な質問やその回答を使用することで，正確な多言語対話が可能となる．

また，我々は用例対訳の収集，共有を目的とした多言語用例対訳共有システム TackPad の開発を行っている [7]．収集した用例対訳は，正確性評価を行った後，多言語対応医療支援システムへの提供を目指している．しかし，本システムでは用例対訳の正確性評価が十分に行えていないという問題を抱えている．本システムでは収集した用例対訳の正確性評価を複数人で行っている．また，専門家である翻訳者の他に，非専門家である一般の利用者にも正確性評価を許可している．これは，翻訳者の絶対数が少ないため，評価者の負担を減らすために行っている．しかし，評価が必要な用例数は多く，十分な評価が得られていない．

そこで本論文では，クラウドソーシング [8], [9] を用いた用例対訳候補の正確性評価手法を提案する．クラウドソーシングとは，人々（群衆）への作業や業務の委託を指す．クラウドソーシングでは安価に大量の用例の評価依頼が可能であるため，不正確な用例対訳の発見を多人数で行うことができる利点がある．ただし，クラウドソーシングで収集したデータは不正確なものが含まれている場合が多く [10]，クラウドソーシングの作業者が行った正確性評価結果をそのまま利用することは難しい．このため，本論文では作業者の作業時間の比を活用し，正確な評価が可能な手法を提案する．

2. 関連研究

本章では，多言語間対話支援技術およびクラウドソーシングを用いたデータ収集に関する関連研究について述べ，本研究との差異を明らかにする．

2.1 多言語間対話支援

多言語間コミュニケーション支援を目的として，用例対訳を用いた支援技術の研究や，機械翻訳を用いた支援技術の研究が多く行われている．機械翻訳は自由に入力された文をすべて多言語に翻訳が可能であるため，子供向けの機械翻訳 [11] や多言語対面環境の討論支援 [12] など，様々な分野で利用されている．しかし，機械翻訳の精度は年々向上しているものの，正確性が求められる医療分野でそのまま利用可能な精度には達していない [13]．また，機械翻訳はルールや統計データに基づいて動的な翻訳を行うため [14]，すべての対訳の正確性を確保することはできない．

そこで現在，正確性が求められる分野においては用例対訳による支援が多く行われている．用例対訳を利用したシステムとして，多言語医療受付支援システム [5] や，携帯型多言語間医療対話支援システム [6] などが存在している．また，用例対訳の収集・共有を目的として，我々は多言語用例対訳共有システム TackPad の開発を行っている [7]．TackPad では，(i) 医療従事者や患者などが必要な用例をシステムに登録，(ii) 翻訳者が (i) で登録された用例を各言語に翻訳，(iii) 作成された用例対訳の正確性評価をシステム利用者が行い，一定の閾値を超えた用例対訳を多言語対応医療システムへ提供する，の手順で，医療現場で使用される用例対訳の収集・共有を Web 上で行っている．

用例対訳の提供には，用例対訳コーパスに含まれているすべての用例，および，すべての言語間の対に対して正確性評価を行う必要がある．現在，TackPad の用例は全言語合わせて約 16,200 文，言語間の対は約 18,500 対，それぞれ存在しているが，正確性評価が不足している．また，医療分野で必要な用例数は 1 言語あたり 3〜5 万文と推測されており [7]，新たな正確性評価手法が必要であると考えられる．

2.2 クラウドソーシングを用いた多言語データ収集

クラウドソーシングを用いたデータ収集は多く行われている．Chen らはクラウドソーシングを用いた類似文の収集 [15] を，Negri らは用例対訳の作成 [16] をそれぞれ行っている．これらの研究では，不適切に対価を得ようとする作業者の存在を考慮した手法が提案されている．また，Callison-Burch はクラウドソーシングを用いて用例対訳の正確性評価を行っている [10]．文献 [10] では作業者の評価の一部と翻訳者の評価との一致率を確認したり，作業者同士の評価を比較したりすることで，不正確な作業者の評価を除去する手法を提案している．しかし，翻訳者を用いる場合は教師データを使用することと同義である．本論文では，教師データを用いずに評価が可能な手法を提案する．

Rzeszotarski らは我々と同様に，教師データを用いずに作業者の行動を利用した正確性評価を行っている [17]．文献 [17] では，作業者の作業時間や作業内容から作業者の質を判定している．また，Harrison らは作業時間が極端に短い作業者を除去している [18] [*1]．このように，作業者の作業時間に着目した正確性評価はいくつか行われている．しかし，文献 [17] の手法ではクラウドソーシングサービスから提供される情報以外に，クリックやスクロールなどの様々な情報を取得する必要がある．本論文では，クラウドソーシングサービスから提供される情報のみで判定を行う．

[*1] その他にも，文献 [18] では，回答の分散値が 0 に近い作業者（異なる設問に同じ回答を続けている作業者）や，同じ内容のタスクに対して大きく異なる回答を行った作業者を除去している．

表 1 使用したデータセットの一例
Table 1 Examples of dataset.

ID	日本語	英語	正確性
58	ポリープをとりたいです	You want to take a polyp	不正確（機械翻訳を使用）
81	日本では，この薬は医師の処方せんがないと入手できません	In Japan, these medicines can be obtained only with a doctor's prescription.	正確
94	ズキンズキンとした痛みがします	It's a throbbing pain.	正確

また，文献 [18] では作業時間の長さを閾値として用いているが，同じ作業でも人によって作業に必要となる時間は異なる．本論文では作業時間の比を用いることで，作業者ごとに異なる，作業に必要な時間の影響の除去を目指す．

3. 正確性評価手法と実験概要

本章では，提案する正確性評価手法と評価実験の概要について述べる．

3.1 作業時間の比を用いた正確性評価手法

本節では，正確な評価を行う作業者を教師データを用いずに選定する手法について述べる．通常の正確性評価では，5 段階評価などの評価結果のみを得ることが一般的である．しかし，評価結果の選択のみを行うタスクの場合，適切な作業を行う作業者と不適切な作業を行う作業者の差が明確になりにくく，分類が困難である場合が多い．

そこで本手法では，適切な作業者と不適切な作業者の作業時間の差が出やすいタスクを用いる．本手法のタスクでは，正確性評価において作業者が不正確な対と判定した場合，正しい翻訳文の入力を作業者に依頼する．反対に作業者が正確な対と判定した場合は，文の入力を依頼しないこととした．不適切な作業者の多くは短い時間で報酬を得るため，翻訳文の入力タスクでは機械翻訳を用いたり，同じ文を入力したりするなどの行動をとる [10], [16]．このような作業者は，文の入力の有無で作業時間がほぼ変わらないと考えられる．反対に，自身の力で翻訳を行う適切な作業者は文の入力の有無で作業時間が大きく変わると考えられる．

これらのことから，本手法では式 (1) を用いて，文の入力あり作業時間と文の入力なし作業時間の比 R を作業者ごとに算出することで作業者の分類を行う．本論文では，比 R が大きい作業者の評価を優先して使用する．

$$R = \frac{\sum_{i=1}^{p} \alpha_i}{p} \Big/ \frac{\sum_{j=1}^{q} \beta_j}{q} \quad (p \geq 2, q \geq 2) \tag{1}$$

α_i は文入力ありの時の作業時間を，β_j は文入力なしの時の作業時間をそれぞれ示す．また，p は文入力ありのタスク数を，q は文入力なしのタスク数をそれぞれ示す．なお，p および q が小さい場合は，個別のタスクの難易度に大きく影響されるため，本論文では $p \geq 2$，$q \geq 2$ となる作業者のデータのみを用いることとする．

表 2 データセットの概要
Table 2 Summary of dataset.

		日本語	英語
単語数	平均	6.7	5.6
	標準偏差	3.1	2.7
構文木の最大の深さ	平均	1.3	3.6
	標準偏差	0.7	2.2

3.2 評価用データセット

本実験では，日英の 100 対（正確 80 対，不正確 20 対）のデータセットを用いた．そのうち 85 対は多言語用例対訳共有システム TackPad [7] 内に存在していた日英の用例対訳，15 対は TackPad 内の日本語用例と，日本語用例を機械翻訳で英語に翻訳した文との対である．

まず，TackPad から 200 対の日英対をランダムに抽出し，日英翻訳者 3 名に評価を依頼した．本論文では 2 言語間の意味比較に用いられる Walker らの適合性評価 [19] を利用して 5 段階評価[*2]を行い，翻訳者評価の平均が 4 未満であった場合，不正確と判定した．この結果，正確が 195 件，不正確が 5 件[*3]となった．このうち，正確と判定された対からランダムに抽出した 80 対と，不正確と判定された 5 対の計 85 対をデータセットに含めることとした．

また，不正確な対を増加させるため，TackPad 内で英語に翻訳されていない日本語用例をランダムに 100 文抽出し，機械翻訳を用いて英語に翻訳した．機械翻訳は言語グリッド [20] を通じて J-Server，WEB-Transer，Google 翻訳を利用した．作成した日英対に対して，Walker らの適合性評価を翻訳者 5〜6 名が行い，平均が 4 未満（不正確と判定）であったもののうち，各機械翻訳の結果からランダムに 5 対ずつ，計 15 対をデータセットに含めた．これら 15 対と先の 85 対とを合わせた計 100 対を本論文ではデータセットとして用いることとした．

データセットの一例を**表 1** に示す．表 1 より，医療従事者の発話内容や患者の発話内容が含まれており，医療分野に特化したデータセットとなっていることが分かる．また，データセットの概要を**表 2** に示す．表 2 のうち，構文木の最大の深さは，構文木のルートからのパスの最大の長

[*2] 評価段階は，1：まったく違う意味，2：雰囲気は残っているがもとの意味は分からない，3：意味はだいたいつかめる，4：文法などに多少問題があるがだいたい同じ意味，5：同じ意味，である．
[*3] 不正確のうち 1 件は，評価者の評価平均では正確と評価されたが，スペルミスが存在していたため最終的に不正確と判定した．

さを用いている．また，日本語は MeCab と CaboCha を，英語は TreeTagger と MaltParser をそれぞれ用いて算出している．

3.3 作業者の正確性評価データの収集

本節では，クラウドソーシングを用いた正確性評価データの収集について述べる．本実験では，クラウドソーシングとして Amazon Mechanical Turk [*4]（以下 AMT とする）を利用した．AMT は文の翻訳作業や画像内の人物の男女判定など，機械にとって難しく人にとっては比較的簡単な作業を，Web 上で安価に依頼が可能なクラウドソーシングサービスである．

本実験では，AMT 上で 3.2 節の日英対（100 対）の評価を依頼した．作業者へは 1 文につき以下の 2 つのタスクを依頼した．

タスク 1 日英対に対して Walker らの適合性評価（5 段階評価）を行う．単語のスペルミスの有無や，文の流暢性評価などを含む，両方の文が同じ意味であるかの確認を作業者に依頼するタスクである．

タスク 2 作業者がタスク 1 で評価を 3 以下とした場合（不正確と評価した場合），翻訳対象文（日本語）を正しい文（英語）へ翻訳するように依頼した．ただし，タスク 1 で 4 以上と評価した場合もタスク 2 での英文入力は可能とした．

なお，1 対につき 10 名の評価を行い，1 対の評価で 5 セントを作業者に支払った[*5]．この結果，34 名の作業者による 997 件の正確性評価結果を取得した[*6]．なお，作業者 1 人あたりの平均評価数は 29 件（最低 1 件，最大 100 件）であった．また，本データ収集は約 12 時間で完了している．

3.4 正確性評価手法の評価方法

本節では，3.3 節で収集した正確性評価結果の分析方法について述べる．本論文では以下の 3 条件について調査を行った．なお，作業時間は，AMT から提供される，「作業者がタスクの結果を提出した日時」と「作業者がタスクに応じた（作業開始した）日時」の差を使用した．

Baseline 任意の作業者 n 名の評価結果の平均

既存手法 極端に短い時間で評価を行った作業者を除去した，任意の作業者 n 名の評価結果の平均

提案手法 作業者を式 (1) の比 R の降順に並び替え，上位 n 名の評価結果の平均

なお，既存手法は文献 [18] を参考に設計を行った．文献 [18] では具体的な作業時間の閾値が記載されていなかったため，本論文で述べる既存手法では，（評価の平均時間 − 標準偏差）を下限値とし，それ以下の作業時間で完了したタスクがあった作業者のデータを除去した．また，本実験では文献 [18] とは異なり，文の入力の有無で作業時間が変化することが予想されるため，文の入力の有無ごとに下限値を設定している．

表 3 各手法で使用した作業者数と評価数
Table 3 Number of evaluations and workers of each method.

	使用データ		未使用データ	
	作業者数	評価数	作業者数	評価数
Baseline	34	997	-	-
既存手法	29	773	5	224
提案手法	13	809	21	188

各手法で使用した作業者数と評価数を**表 3** に示す．表 3 より，既存手法は作業者 5 名の 224 件の評価結果を使用していないことが分かる．また，提案手法は作業者 21 名の 188 件の評価結果を使用しておらず，主に評価量が少ない作業者のデータが使用されていないことが分かる．

また，本分析は次の手順で行った．本分析では使用する評価数 n を 1〜5 に変化させて，各評価手法の評価を行った．提案手法は適切な作業者の順序づけを行う手法であるため，順位が高いものを優先する必要がある．このため，既存手法や Baseline も含めて，評価数を制限して比較を行っている．

(1) 前述の各手法で取得した評価結果の平均値が 4 以上の場合は正確，4 未満の場合は不正確と判定する．その上で 3.2 節のデータセットと評価が一致した場合は正解，一致しない場合は不正解とする．

(2) 計 100 対に対して (1) の操作を行い，正解率を取得する．提案手法以外は 10 回試行した平均値を用いる[*7]．

また，次章の考察では，比較対象として任意の翻訳者 1 名の評価結果を用いる．本評価についても上記の手順で 10 回の試行を行った平均を用いる．

4. 分析と考察

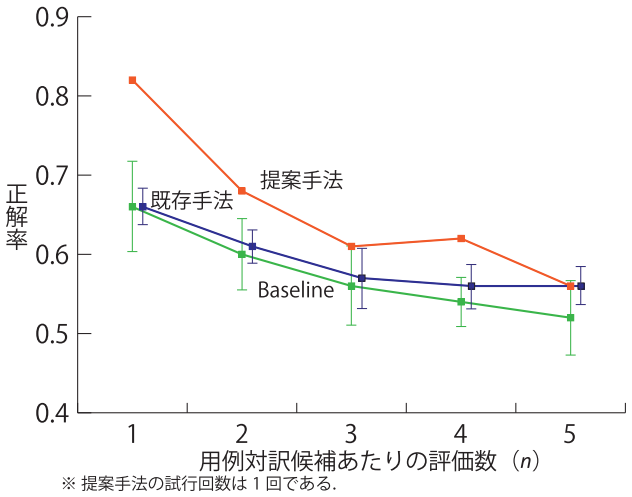
本章では前章で述べた各手法の分析と考察を行う．

4.1 各手法の判定結果と考察

本論文では主に各手法の正解率とカッパ係数をそれぞれ用いて議論を行う．カッパ係数は，偶然による一致の影響を取り除いた指標であり，−1〜1 の値となる．カッパ係数が 1 のときは完全一致，0 のときは無関連をそれぞれ示している．本論文で扱う正解データは，8 割が正確，2 割が不正確となるデータである．すべて正確と評価すると 8 割の正解率となるため，カッパ係数を併用して議論する．なお，任意の翻訳者 1 名の評価の場合，正解率が平均 90.0%，正解データとのカッパ係数が平均 0.62 であった．

各手法の正解率を**図 1** に示す．図 1 より，各用例対訳

[*4] https://www.mturk.com/
[*5] AMT では比較的高い支払額である．
[*6] 1,000 件中 3 件にデータの欠落が存在していたためである．
[*7] 提案手法以外は任意の評価値を取得しており，試行ごとに評価値が変化するため．



※ 提案手法の試行回数は 1 回である.

図 1 各手法の正解率
Fig. 1 Accuracy rates of each method.

表 4 正解データと各手法とのカッパ係数
Table 4 Kappa statistics of correct data and respective methods.

用例対訳候補あたりの評価数（n）	1	2	3	4	5
Baseline	0.10	0.09	0.09	0.08	0.07
既存手法	0.07	0.09	0.09	0.10	0.11
提案手法	0.44	0.25	0.17	0.18	0.12

・Baseline および既存手法は 10 回試行したものの平均である.

候補あたりの評価数 n が 1 のとき，提案手法の正解率は 82.0%であり，Baseline および既存手法よりも約 16 ポイント，正解率が向上していることが分かる[*8].

正解データと各手法とのカッパ係数を**表 4** に示す．表 4 より，Baseline および既存手法のカッパ係数は最大 0.11 であり，信頼性が低いことが分かる．提案手法は n の値が小さいときに他の手法と比較して高いカッパ係数の値を示しており，$n = 1$ のときは 0.44 と中程度の一致となっている．

これらのことから，作業時間の比を用いた提案手法は，n の値が低い場合において，Baseline や既存手法よりも正確な正確性判定が可能であることが分かる．なお，既存手法は $n = 5$ の際に提案手法と同程度の正解率となったが，両手法ともカッパ係数の値が低いため，十分な信頼性が担保できていないと考えられる．これは，n が大きくなると，比 R が低い作業者の評価結果を提案手法でも使用することが理由として考えられる．

4.2 作業時間の比を用いた効果

本節では，提案手法で用いた作業時間の比の効果の確認を目的として，作業者ごとのデータを用いて考察を行う．

[*8] 既存手法のもととなった文献 [18] では，他の指標も存在していた．本実験に適用可能な，回答の分散値が 0 に近い作業者（具体的な閾値が記述されていなかったため，分散が 0 かつ 3 件以上評価した作業者とした）を除去した結果，1 名の 3 件の評価が削除されたのみで，結果に大きな影響を与えなかった．

表 5 作業者の作業時間と比 R
Table 5 Working time and rate R of workers.

	作業時間（秒）		比 R
	入力あり	入力なし	
平均	33.9	24.0	2.0
標準偏差	19.2	20.8	1.2
最大値	141	172	5.7
最小値	7	3	1.0

提案手法で用いた作業者の作業時間と比 R の結果を**表 5** に示す．表 5 より，文の入力ありの方がタスク完了までの時間が長い傾向があることが分かる．また，比 R は 1.0〜5.7 まで広い値を取っている．なお，比 R の平均値である 2.0 を越えたのは，評価者 13 名中 3 名のみであった．

表 6 に提案手法における，作業者の人数別の正解率を示す．表 6 では，例として上位 3 名の作業者の評価（m）を 2 名分使用（n）した場合，27 件の用例対訳候補の評価を行い，正解率が 85.2%，カッパ係数が 0.57 であったことを示している．なお，図 1 の提案手法の正解率は，表 6 中の評価数が 100 件のときの正解率にあたる．表 6 では $m = 8$ 以降を省略しているため，図 1 の正解率と対応した値は $n = 2$ まで示されている．

表 6 より，評価に利用する作業者数 m を絞ると正解率，カッパ係数ともに高くなり，上位 4 名までの評価は正解率が 80%以上，カッパ係数も 0.4 以上と高い値となっていることが分かる．また，$n = 2$，$m = 2$ や $n = 4$，$m = 6$ など，利用する作業者数 m を制限した上で評価数 n を増やすと，高い正解率およびカッパ係数になることが分かる．特に，$m = 1$ のときは正解率が 95.8%，カッパ係数が 0.86 であり，4.1 節で述べた翻訳者 1 名の値よりも高くなっている．これらは，作業時間の比 R が大きい作業者から優先選択する提案手法の優位性が示されている．

なお，AMT では不正確な評価を行う作業者に対価を支払わないことが可能である．このため，実際の運用では本手法を用いて適切な評価結果のみを得ることで，さらに高い正解率となる評価結果を得ることができると考えられる．

4.3 不正確な用例対訳候補の評価結果

本節では，データセット中の不正確な用例対訳候補の評価結果の詳細について述べ，考察を行う．

正確性評価では，翻訳者の方がクラウドソーシング上の作業者よりも正確性が高い傾向にある．しかし，一部の対の評価では，作業者の方が正確に評価できた例も存在していた．**表 7** に作業者の方が正確に評価できた対の作業者と翻訳者の評価例を示す．なお，表中の翻訳者の列は，同一人物の評価である．また，各用例対訳の 10 名の作業者は同一ではないため，作業者の各列は別の作業者の評価結果を示している．また，作業者の評価は左側から昇順に並べ

表 6 作業者の人数別正解率（提案手法）
Table 6 Accuracy rates by number of workers (proposed method).

利用した作業者の人数（m）	用例対訳候補あたりの評価数（n）				
	1	2	3	4	5
1	95.8% (48) ⟨0.86⟩				
2	86.9% (84) ⟨0.57⟩	88.0% (25) ⟨0.65⟩			
3	87.2% (86) ⟨0.57⟩	85.2% (27) ⟨0.57⟩	100.0% (2) ⟨–⟩		
4	82.8% (93) ⟨0.47⟩	85.1% (47) ⟨0.50⟩	100.0% (6) ⟨1.00⟩	-	
5	81.4% (97) ⟨0.43⟩	75.3% (73) ⟨0.38⟩	80.0% (20) ⟨0.41⟩	100.0% (1) ⟨–⟩	-
6	82.0% (100) ⟨0.44⟩	69.1% (97) ⟨0.27⟩	67.1% (73) ⟨0.27⟩	90.0% (20) ⟨0.62⟩	100.0% (1) ⟨–⟩
7	82.0% (100) ⟨0.44⟩	68.0% (100) ⟨0.25⟩	60.8% (97) ⟨0.17⟩	67.1% (73) ⟨0.27⟩	70.0% (20) ⟨0.29⟩

・表中の値は，順に正解率，評価数，カッパ係数である．
・評価数は，上位 m 名の作業者の評価のうち，n 名分使用するとしたときに，何件の評価が可能であったかを示す．

表 7 作業者と翻訳者の評価
Table 7 Selection of evaluation results of workers and translators.

ID	日本語	英語	間違いの理由	翻訳者			作業者								
				A	B	C									
28	肺炎	neumonia	スペルミス (正しくは “Pneumonia”)	1	5	4	1	2	3	4	4	5	5	5	5
31	放射線腫瘍科	external beam radiotherapy	意味が違う (英語は「対外照射療法」の意)	5	2	4	1	1	1	2	2	2	4	4	5
85	子宮がん	uterrine cancer	スペルミス (正しくは “uterine cancer”)	5	4	5	2	4	4	5	5	5	5	5	5

・翻訳者，作業者の各列の数字は，5 段階評価結果をそれぞれ示す．
・翻訳者の列は，同じ翻訳者（翻訳者 A～C）の評価である．作業者の評価結果は昇順に並べ替えを行っている．

替えを行っている．

表 7 の ID85 は，英文にスペルミスが含まれていた．しかし，3 名の翻訳者はスペルミスを発見できていないことが分かる．10 名中 9 名の作業者も評価 4 以上をつけていたが，1 名の作業者は評価 2 をつけ，タスク 2（正しい英文の記入）で正しい英単語を記述していた．ID28 も同様に，翻訳者のうち 1 名のみがスペルミスを発見したが，作業者は 3 名がスペルミスを発見していた．また，ID31 は意味が異なる用例が対として表示されていた．表 7 より，翻訳者は 3 名中 1 名のみ誤りを指摘していたが，作業者は 10 名中 6 名が誤りを指摘していることが分かる．なお，誤りを指摘した翻訳者は用例対訳候補によって異なっていることから，翻訳者らの能力に大きな優劣はないと考えられる．

これらの結果から，一部の対に対する作業者の評価は，翻訳者の評価と同等もしくはそれ以上に適切な評価が行われていることが分かる．このため，作業者の評価を含めることにより，翻訳者が発見できなかった不正確な文を発見可能になる利点が存在すると考えられる．本論文では，作業者同士の作業結果の一致率を用いた文献 [10] との比較を行っていないが，文献 [10] の手法では，前述したような一部の作業者のみが正しい評価を行った場合は正しい評価ができない可能性が高いと考えられる．

なお，ID28 や ID85 のように，少数の作業者が不正確と判定した場合，提案手法を用いた場合でも作業時間の比によっては抽出できない可能性があると考えられる．このため，用例対訳候補が不正確であると評価した評価者がいた場合は，要注意の用例対訳候補であると扱うなど，別手法との併用が必要になると考えられる．

5. おわりに

本論文では，正確な多言語間コミュニケーションに用いる用例対訳の正確性評価手法について述べた．本手法では，異なる作業時間となるように設計されたタスクをクラウドソーシング上の作業者に依頼することで，教師データを用いずに正確性評価を行っている．本研究の貢献は，タスク作業時間の比をもとにした作業者の順位付けを行うことにより，従来手法よりも高い正確性を持つ正確性評価手法を提案した点である．

今後は，必要な評価数 n や，利用する作業者の人数 m，比 R の値などを用いた適切な閾値の調査を行う．また，用例対訳の提供時に，専門家である翻訳者の評価が含まれていないことによる利用者の不安を軽減するため，少数の翻訳者の評価を併用することも今後の検討事項である．

謝辞 本研究の一部は，JSPS 科研費 JP22300044, JP26730105 による．

参考文献

[1] 総務省：多文化共生の推進に関する研究会報告書，総務省（オンライン），入手先 ⟨http://www.soumu.go.jp/kokusai/pdf/sonota_b5.pdf⟩（参照 2017-04-04）．
[2] Takano, Y. and Noda, A.: A temporary decline of thinking ability during foreign language processing, *Journal of Cross-Cultural Psychology*, Vol.24, pp.445–462 (1993).
[3] Aiken, M., Hwang, C., Paolillo, J. and Lu, L.: A group

decision support system for the Asian Pacific rim, *Journal of International Information Management*, Vol.3, No.2, pp.1–13 (1994).
[4] Kim, K.J. and Bonk, C.J.: Cross-Cultural Comparisons of Online Collaboration, *Journal of Computer Mediated Communication*, Vol.8, No.1 (2002).
[5] 宮部真衣，吉野　孝，重野亜久里：外国人患者のための用例対訳を用いた多言語医療受付支援システムの構築，電子情報通信学会論文誌，Vol.J92-D, No.6, pp.708–718 (2009).
[6] 尾崎　俊，松延拓生，吉野　孝，重野亜久里：携帯型多言語間医療対話支援システムの開発と評価，電子情報通信学会技術研究報告，Vol.AI2010-47, pp.19–24 (2011).
[7] 福島　拓，吉野　孝，重野亜久里：正確な情報共有のための多言語用例対訳共有システム，情報処理学会論文誌コンシューマ・デバイス&システム，Vol.2, No.3, pp.23–33 (2012).
[8] Howe, J.: Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business, Crown Business (2008).
[9] Doan, A., Ramakrishnan, R. and Halevy, A.Y.: Crowdsourcing systems on the World-Wide Web, *Comm. ACM*, Vol.54, No.4, pp.86–96 (2011).
[10] Callison-Burch, C.: Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk, *Proc. 2009 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2009*), pp.286–295 (2009).
[11] Matsuda, M. and Kitamura, Y.: Development of Machine Translation System for Japanese Children, *Proc. 2009 ACM International Workshop on Intercultural Collaboration* (*IWIC'09*), pp.269–271 (2009).
[12] 福島　拓，吉野　孝，喜多千草：共通言語を用いた対面型会議における非母語話者支援システム PaneLive の構築，電子情報通信学会論文誌，Vol.J92-D, No.6, pp.719–728 (2009).
[13] 林田尚子，石田　亨：翻訳エージェントによる自己主導型リペア支援の性能予測，電子情報通信学会論文誌，Vol.J88-D1, No.9, pp.1459–1466 (2005).
[14] 塚田　元，渡辺太郎，鈴木　潤，永田昌明，磯崎秀樹：統計的機械翻訳，NTT 技術ジャーナル，Vol.19, No.6, pp.23–25 (2007).
[15] Chen, D.L. and Dolan, W.B.: Collecting Highly Parallel Data for Paraphrase Evaluation, *Proc. 49th Annual Meeting of the Association for Computational Linguistics*, pp.190–200 (2011).
[16] Negri, M. and Mehdad, Y.: Creating a Bi-lingual Entailment Corpus through Translations with Mechanical Turk: $100 for a 10-day Rush, *Proc. NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp.212–216 (2010).
[17] Rzeszotarski, J.M. and Kittur, A.: Instrumenting the crowd: Using implicit behavioral measures to predict task performance, *Proc. 24th Annual ACM Symposium on User Interface Software and Technology* (*UIST 2011*), pp.13–22 (2011).
[18] Harrison, C., Horstman, J., Hsieh, G. and Hudson, S.E.: Unlocking the Expressivity of Point Lights, *Proc. SIGCHI Conference on Human Factors in Computing Systems* (*CHI 2012*), pp.1683–1692 (2012).
[19] Walker, K., Bamba, M., Miller, D., Ma, X., Cieri, C. and Doddington, G.: Multiple-Translation Arabic (MTA) Part 1, *Linguistic Data Consortium, Philadelphia* (2003).
[20] Ishida, T. (Ed.): *The Language Grid: Service-Oriented Collective Intelligence for Language Resource Interoperability*, Springer (2011).

福島 拓 （正会員）

1986 年生．2008 年和歌山大学システム工学部中退．2013 年同大学大学院システム工学研究科博士後期課程修了．博士（工学）．現在，大阪工業大学情報科学部特任講師．CSCW の研究に従事．

吉野 孝 （正会員）

1969 年生．1992 年鹿児島大学工学部卒業．1994 年同大学大学院工学研究科修士課程修了．博士（情報科学）．現在，和歌山大学システム工学部教授．CSCW，HCI の研究に従事．