クラウドソーシング労働者の作業特徴に着目した多言語テキストペアの正確性評価手法

福島 拓[1,a)]　吉野 孝[2,b)]

概要：現在，グローバル化による多言語間コミュニケーションの機会が増加している．しかし，多言語間での正確な情報共有は十分に行われていない．正確な多言語支援が求められる場では，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，正確性評価を十分に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，Web 上での不特定多数の人への作業委託を行うことが可能な，クラウドソーシングを用いた多言語テキストペア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である Turker の中で，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．本稿の貢献は次の 2 点である．(1) タスク作業時間を元にした Turker の優先順位付けを行うことにより，翻訳者に準ずる正確性評価を可能にする手法を提案した．(2) 一部の多言語テキストペアにおいては，正確性の高い評価を行う翻訳者と同等もしくは高い精度で Turker が評価を行えていることを示した．

Correctness Evaluation Method of Multilingual Texts Pairs of Crowdsorcings' Workers

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

Abstract: Recently, worldwide globalization has helped to increase communication among people with different native languages. However, it is not enough that multilingual accurate information sharing. In general, multilingual support systems for applications that require high accuracy use multilingual texts pairs called parallel texts. Accuracy evaluating of parallel texts is important for providing of parallel texts. Since there are huge quantities of parallel texts, they become difficult to fully collect accuracy evaluations. Therefore, we propose an correctness evaluation method of multilingual texts pairs that used crowd sourcing. Crowd sourcing is the service which can perform work commission to many and unspecified persons on Web. Our proposed method can choose preferentially a Turker, which is a worker on crowd sourcing, evaluated correctly. The contributions of this paper are the following two items. (1) We proposed an correctness evaluation method of multilingual texts pairs that used crowd sourcing. The idea of this method is chosen preferentially a Turker based on task working time. This method enables accuracy evaluation according to a translator. (2) From the result of our experiment, our proposed method can obtain high accuracy as same as a human translator who performs evaluation with high accuracy.

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．日本国内でも在日外国人数や留学生数，訪日外国人数は 10 年前のそれぞれ約 1.3 倍，約 1.4 倍，約 1.2 倍と増加傾向にあり [1], [2], [3]，今後，外国人住民のさらなる増加が予想されている [4]．この

[1] 和歌山大学大学院システム工学研究科
Graduate School of Systems Engineering, Wakayama University, Wakayama 640–8510, 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)] fukushima@yoslab.net
[b)] yoshino@sys.wakayama-u.ac.jp

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 [4]，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 [5], [6], [7]，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取りは，意思の疎通を十分に行うことができない．現在，日本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．また，通訳者の身分保障や通訳者自身のメンタルケアなどの問題が存在している [8]．

このような問題は，外国人が多くない地域でも対応する必要性が出てきている．2007 年度の外国人登録者数が全国 22 位 [1] の宮城県において行われた調査 [9] では，79%の医療機関が日本語の不自由な外国人の対応を行っている．内訳としては，中国人患者が 64%，韓国人患者が 34%などとなっている．しかし，外国語対応体制がある医療機関は 36%にとどまっている．また，内訳を見ると中国語が 20%，韓国朝鮮語が 12%などとなっており，対応可能言語と患者の母語が一致していない．このように，医療機関を訪れる外国人患者支援は十分であるとはいえない．

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 [10], [11], [12], [13]．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テキストペアのコーパスのことを指し[*1]，「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(タックパッド) の開発を行っている [14]．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．

そこで本稿では，クラウドソーシング [15], [16] を用いた多言語テキストペアの正確性評価を検討する．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大量の用例に対して安価で評価依頼を行うことができるため，不正確な文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 [17]，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，クラウドソーシングで収集したデータを用いて，不正確なデータの影響が少ない正確性評価手法の検討を行う．

2. 関連研究

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 [18] や多言語対面環境の討論支援 [19] 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 [20]．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 [21]，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 M^3（エムキューブ）[10] や，ケータイ多言語対話システム [11] がある．M^3 はタッチパネルで操作可能としたシステムで，対話機能，外国人患者の受診支援機能 (問診機能，受診科選択機能など) を有している．また，ケータイ多言語対話システムは多言語問診を携帯電話上で実現している．また，自由文に対応するために用例対訳と機械翻訳を併用したシステムも提案されている [12], [13]．

このような用例対訳の収集・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 [14]．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) システム利用者が作成された用例対訳の正確性評価を行い，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，医療現場で求められている用例対訳の収集・共有を Web 上で行っている．

本システムで収集対象としている用例対訳は多くの用例と言語間の対があり，これらすべての評価をシステム上で行う必要がある．現在，本システム上には用例数は全言語合わせて約 14,000 文，言語間の対は約 18,000 対が存在している．また，本システム上の用例対訳の数は不十分であ

[*1] 本稿では，正確性の確保が行われた多言語の対のことを「用例対訳」，正確性の確保が行われていない多言語の対のことを「多言語テキストペア」とする．

ることが分かっている [22]．今後，医療分野で必要な用例対訳を網羅した場合，現在の約数十倍の用例が必要であると考えられる．しかし，現在でも正確性評価が不足しており，新たな方法での正確性評価を得る必要があると考えられる．

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 [23] を，Negri らはクラウドソーシングを用いた多言語テキストペアの作成 [24] 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クラウドソーシング上の作業者が存在すること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多言語テキストペアの正確性評価を行っている [17]．Callison-Burch はクラウドソーシングでの評価と翻訳者の評価との相関関係を確認したり，クラウドソーシング上の作業者同士の評価を比較したりすることで，不正確な作業者の情報の除去する手法の提案をいくつか行っている．しかし，手法の一つはクラウドソーシングの作業者の評価を翻訳者をもとに選択しているため，あらかじめ翻訳者の評価が必要となる．また，翻訳者の評価を全て正しいものとして用いているが，翻訳者の評価がどれだけ適切なものかを評価していない．本稿では，クラウドソーシングの労働者と翻訳者それぞれの評価の適切性について検証した後，さらに適切な正確性評価手法の検討を行う．その際，事前に翻訳者の評価を用意したり，Turker 同士の比較を行わない手法を提案する．

3. クラウドソーシングを用いた正確性評価

3.1 評価用データセット

本節では，評価用のデータセットについて述べる．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 TackPad[14] 内に存在していた日英の用例対訳，15 対は TackPad 内の日本語用例を機械翻訳で英語に翻訳した文である．

TackPad 内の 85 対の日英の用例対訳は日英の翻訳者 3 名に評価を依頼した．本稿では 2 言語間の意味比較に用いられる Walker らの適合性評価 [25] を利用して 5 段階評価を行った．評価基準は，2 言語間の意味が「1:None, 2:Little, 3:Much, 4:Most, 5:All」のいずれに当たるかである．本稿では，評価者の評価の平均が 4 以下だった場合，不正確と判定した．4 は「Most(文法などに多少問題があるがだいたい同じ意味)」であるため，厳しめに判定を行っている．この基準で用例対訳の正確性評価を行った結果，正確が 80 件，不正確が 5 件[*2]となった．

また，日本語用例を機械翻訳で英語に訳した 15 対は，Walker らの適合性評価を翻訳者 5〜6 名が行い，平均が 4 以下 (不正確と判定) であったものを利用している．機械翻訳は J-Server，WEB-Transer，Google 翻訳を利用し，各機械翻訳の結果から 5 対ずつ用いている．

なお，これらの評価では，1 対の評価につき 15 円を翻訳者に支払っている[*3]．

3.2 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 MTurk とする）を利用した．MTurk 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，MTurk 上で 3.1 節の 100 対の日英対評価を依頼した．以下，MTurk 上の作業者を Turker と呼ぶ．Turker へは以下の二つのタスクを依頼した．

タスク 1　前節と同様に Walker らの適合性評価 (5 段階評価) を行う．

タスク 2　Turker 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)，日本語を正しい英文に翻訳．

なお，1 対につき 10 名の評価を行い，1 対の評価で 0.05 ドルを Turker に支払った[*5]．この結果，34 名の Turker による 997 件の正確性評価結果を取得した[*6]．なお，Turker 一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，これらのタスクは約 12 時間で完了している．

4. 分析と考察

4.1 正確性評価の分析方法

本節では，3.2 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．本分析では，Turker の人数を変化させて評価手法の検討を行う．

本分析は次の手順で行った．

(1) Turker の評価結果のうち n 名のデータを取得する．

(2) 翻訳者の評価結果のうち m 名のデータを取得する．

(3) (1) と (2) を合わせた結果の平均を取得する．平均値が 4 より大きい場合は正確，4 以下の場合は不正確と判定し，3.1 節のデータセットの正確・不正確と一致しているかを調べる．データセットと評価が一致した

[*2] 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ている．

[*3] 比較的安価である．エキサイト翻訳依頼プロ (http://www.excite.co.jp/world/order/pro/) の場合，日本語 1 語あたり 25 円が必要なため (2012/09/12 時点)，1 対あたり約 300 円が翻訳に必要となる (本データセットの日本語の平均文数は 11.9 語である)．

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

[*5] 約 3.9 円 (2012/9/12 時点)．MTruk 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．

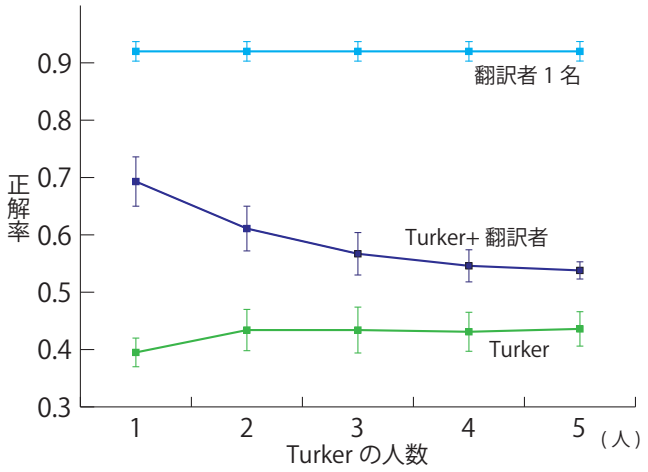


図 1 正確性評価の分析結果
Fig. 1 Analysis result of correctness evaluation.

場合は正解，一致しない場合は不正解とする．

(4) (1)〜(3) の操作を各対に対して行い，Turker n 人，翻訳者 m 人の際の正解率を取得する．

(5) (4) の操作を (1) と (2) で取得するデータをランダムに変えながら 10 回行い，それらの平均から最終的な正解率を算出する．

なお，Turker の人数 n は 0〜5 名を用いた．また，翻訳者の人数 m は 0〜1 人を用いた．

4.2 分析結果と考察

4.2.1 Turker と翻訳者の正解率

本項では，Turker と翻訳者それぞれの正確性評価の正解率について考察する．

分析結果を**図 1** に示す．図 1 より，翻訳者 1 名のみでの評価を行った場合の正解率の平均は 92%であった．それに対し，Turker の評価に翻訳者の評価を含めない場合の正解率は 5 割以下であったことが分かる．これは，一部の Turker が正確に評価を行っていなかったことが原因であると考えられる．この結果から，Turker の正確性評価の結果をそのまま利用することは難しいと考えられる．

ただし，すべての Turker の評価が不正確ではなかった．**表 1** に 5 件以上を評価した 20 名の Turker と正解データセットとの相関係数を示す．表 1 より，相関係数が高い Turker が存在していることが分かる．これらの Turker は翻訳者に準ずる評価精度であるため，これらの Turker の評価を優先的に選択する必要があると考えられる．ただし，実利用時では正解データは存在しないため，正解データとの相関係数以外の評価軸を用いて評価精度の高い Turker を選択する必要がある．

次に，Turker と翻訳者 1 名の評価結果を合わせた結果について考察する．図 1 より，Turker の評価に翻訳者 1 名の評価を入れた場合，約 10〜30 ポイント，正解率が増加していることが分かる．このため，一般の利用者の評価に翻訳者の評価を入れることで，正確性が大きく向上することが分かる．

表 1 データセットと Turker の相関
Table 1 Correlation of evaluation result between the data set and Turkers.

相関係数	〜0	0-0.2	0.2-0.4	0.4-0.6	0.6〜	合計
Turker	3	9	5	2	1	20

・単位は人である．

これらから，Turker の中から評価精度の高い Turker を選択し，翻訳者の評価と合わせることで正確性の高い評価が可能であると考えられる．

4.2.2 不正確な評価データの評価結果

本項では，評価データ中の不正確なデータが各被験者にどのように評価されたかについて述べ，考察を行う．

正確性評価の正解率は前項で述べたとおり，翻訳者の方が Turker よりも高い正解率であった．しかし，一部の対の評価は，Turker の方が正確に評価できた例も存在していた．**表 2** に Turker 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 Turker は同一ではないため，Turker の各列は別の Turker を示す．また，Turker の評価は左側から昇順に並び替えを行っている．

表 2 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 Turker も評価 4 以上をつけていたが，1 名の Turker 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，Turker は 3 名がスペルミスを発見していた．また，ID31 は意味が異なる日英文が対として表示されていた．表 2 より，翻訳者は 3 名中 1 名のみが間違いを指摘していたが，Turker は 10 名中 6 名が間違いを指摘していることがわかる．なお，間違いを指摘した翻訳者はテキストペア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，翻訳者は全ての間違いを発見できていないことが分かる．このため，翻訳者 1 名のみで評価を行うと，不正確な対を見落とす可能性があることが分かる．また，一部の対に対する Turker 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．これらのことから，Turker 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できる可能性が考えられる．

5. 良質な評価者の抽出

5.1 良質な評価者の優先選択

前章での分析より，非専門家である Turker は専門家である翻訳者より，平均的には翻訳精度が低いという結果と

表 2　Turker と翻訳者の評価の一部

Table 2　A part of evaluation result of Turkers and translators.

ID	日本語	英語	間違いの理由	翻訳者			Turker								
				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	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	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	5

・翻訳者，Turker の各列の数字は，それぞれの評価者の 5 段階評価を示す．
・翻訳者の列は，同じ翻訳者 (翻訳者 A〜C) の評価である．Turker の評価結果は昇順に並び替えを行っている．

表 3　文の入力の有無による Turker の作業時間

Table 3　Working time of with or without inputting text of Turkers.

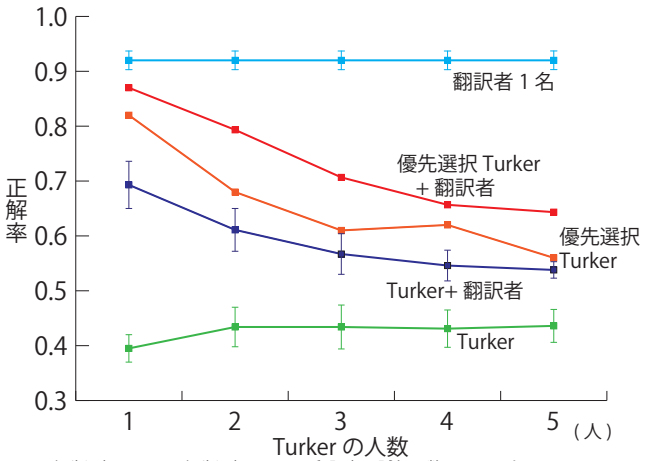
Turker	文の入力時間の平均		割合	相関係数	評価数
	あり (α)	なし (β)	(α/β)		
1	26.9	4.7	5.7	0.72	48
2	36.1	12.3	2.9	0.30	61
3	59.5	22.8	2.6	0.37	6
4	28.0	14.1	2.0	0.32	31
5	34.5	20.1	1.7	0.01	45
6	38.2	23.5	1.6	0.21	100
7	36.0	24.0	1.5	0.13	100
8	28.5	20.8	1.4	0.49	8
9	32.9	24.3	1.4	0.16	79
10	32.5	24.3	1.3	0.11	99
11	21.0	15.7	1.3	0.20	37
12	36.8	29.5	1.3	0.20	100
13	39.2	39.1	1.0	0.06	95

・文の入力あり，文の入力なしの単位は秒である．
・表中の相関係数は Turker 評価と正解データとの相関係数を示す．
・表中の評価数は，各 Turker が評価したテキストペアの数を示す．
・α/β の値が大きい Turker から順に表示している．

なった．しかし，翻訳者が発見できなかった不正確な対を発見できた Turker も存在している．そこで本節では，不正確な評価を行う Turker を収集データから除き，正確な評価を行う良質な Turker を抽出する手法の検討を行う．

不正確な評価を行う Turker は，正確な英文の記入を求めた「タスク 2」において，機械翻訳を利用している例が多く見られた．機械翻訳は，人手で行う日英翻訳よりも早く行うことができると考えられる．ただし，1 つのタスクにかける時間は Turker によって異なる．このため，文の入力の有無 (タスク 2 を行ったかどうか) による作業時間の変化を Turker ごとに調査した．

表 3 に文の入力の有無による Turker の作業時間，および各 Turker の評価と正解データとの相関係数を示す．表中の α は文の入力ありの時にタスク終了までにかかった時間の平均を，β は文の入力なしの時にタスク終了までにかかった時間の平均をそれぞれ示す．なお，表 3 は文の入力ありとなしのそれぞれを 2 件以上行った利用者 13 名のみ示している．また，文の入力ありとなしの差が大きかったものから順に示している．表 3 中の α/β の値が大きい



※ 優先選択 Turker，優先選択 Turker+ 翻訳者の試行回数は 1 回である．

図 2　正確性評価の分析結果 (Turker の優先順位利用)

Fig. 2　Analysis result of correctness evaluations (used prioritized Turkers).

Turker は文の入力に時間をかけている Turker であると考えられる．逆に，α/β の値が小さい Turker は文の入力に時間をかけていない，つまり，機械翻訳を利用している可能性が高い Turker であると考えられる．

表 3 より，α/β の値が大きい Turker の評価結果は，比較的正解データとの相関係数が大きい傾向にあることが分かる．このことから，Turker の文の入力の有無による作業時間の割合を利用することで，正確な評価を行う Turker が抽出できる可能性があることが分かる．

この抽出手法を用いて，4.1 節の評価を再度行った．優先順位付けに使用する割合を求める式を，式 (1) に示す．

$$R = \frac{\sum_{i=1}^{n} \alpha_i}{n} \Big/ \frac{\sum_{j=1}^{m} \beta_j}{m} \quad (n \geq 2, m \geq 2) \qquad (1)$$

式 (1) の α_i は文の入力ありの時にタスク終了までにかかった時間を，β_j は文の入力なしの時にタスク終了までにかかった時間をそれぞれ示す．また，n，m は，それぞれ α_i と β_j の数を示す．本章では，式 (1) の割合 R を用いて，4.1 節の正解率算出方法の (1) でランダムに取得していた Turker の評価を，R が大きい Turker の評価を優先して使用するように変更した．

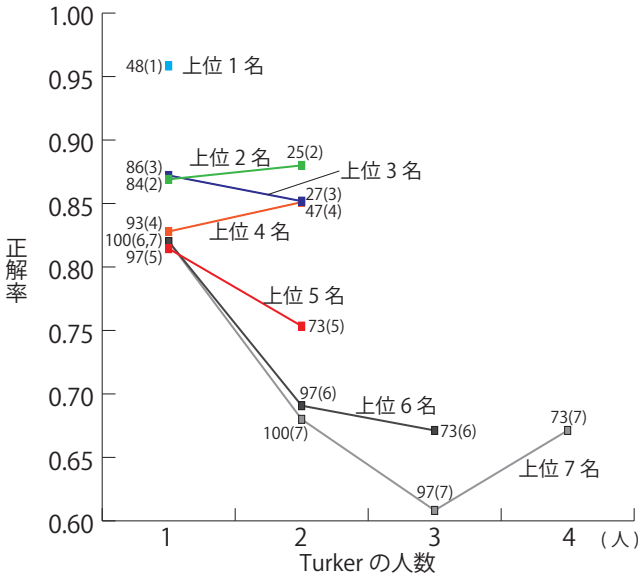
式 (1) を元にした Turker の優先順位付けを行った場合の正解率を**図 2** に示す．図 2 より，Turker を 1 名のみ利

表 4　優先順位を用いた Turker の人数別正解率

Table 4　The accuracy rate according to number of the prioritized Turkers.

利用した Turker の人数		テキストペアあたりの評価人数				
		1	2	3	4	5
1	正解率	95.8%	-	-	-	-
	評価数	48	0	0	0	0
2	正解率	86.9%	88.0%	-	-	-
	評価数	84	25	0	0	0
3	正解率	87.2%	85.2%	100%	-	-
	評価数	86	27	2	0	0
4	正解率	82.8%	85.1%	100%	-	-
	評価数	93	47	6	0	0
5	正解率	81.4%	75.3%	80.0%	100%	-
	評価数	97	73	20	1	0
6	正解率	82.0%	69.1%	67.1%	90.0%	100%
	評価数	100	97	73	20	1
7	正解率	82.0%	68.0%	60.8%	67.1%	70.0%
	評価数	100	100	97	73	20

・表中の評価数は，上位 n 名の Turker の評価を用いたときに，何件の評価を行ったかを示す．



※グラフ中の数字は評価対象テキストペアの数を，括弧内は上位何人の Turker を利用したかをそれぞれ示す．

図 3　優先順位を用いた Turker の人数別正解率 (20 件以下の評価数を除く)

Fig. 3　The accuracy rate according to number of the prioritized Turkers (except for 20 or less numbers of evaluations).

用する場合，Turker の優先順位付けを利用するとランダムに選択したときより 42 ポイント正解率が向上したことが分かる．また，翻訳者 1 名と Turker 1 名の評価を合わせて利用する場合，Turker の優先順位付けを利用すると 19 ポイント正解率が向上したことが分かる．特に，翻訳者 1 名と優先順位付けを利用した Turker 1 名の評価を合わせた結果は，翻訳者 1 名の結果結果に近い正解率となったことが分かる．

これらの結果から，潜在的な情報である文の入力時間を元にした Turker の選択により，効果的に良質な評価者が選択できていることが分かる．ただし，Turker の評価の数を 5 名にまで増加させた場合は従来のランダム選択との差が小さくなる．これは選択すべき人数が増えることにより優先順位付けの効果が小さくなるためであると考えられる．

このため，Turker の人数を変化させたときに，正解率がどのように変化するかの調査を行った．**表 4** に Turker の人数を変化させた時の正解率を示す．例えば，優先順位付けを行った Turker の上位 4 名のデータのみを使用し，テキストペアの評価を 1 人だけで行ったときの正解率を求めた場合，93 件のテキストペアが評価可能[*7]で，82.8% の正解率となることが表 4 から分かる．また，表 4 のうち評価数が 20 件以下[*8]のものを除いたデータを元にグラフ化したものを**図 3** に示す．表 4 および図 3 より，Turker の上位 4 名までの評価のみを使用した場合，テキストペアあたりの評価人数を増やしても正解率がほぼ横ばいとなる傾向にあることが分かる．これは，Turker の利用人数の制限を行っていない図 2 の傾向とは異なっている．このため，評価に使用する Turker を優先順位付けした後，利用する Turker の人数を制限することで精度よく正確性評価を行うことが可能であると考えられる．

なお，MTurk では不正確な評価を行う Turker 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の仕組みをもとに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データを得ることができると考えられる．

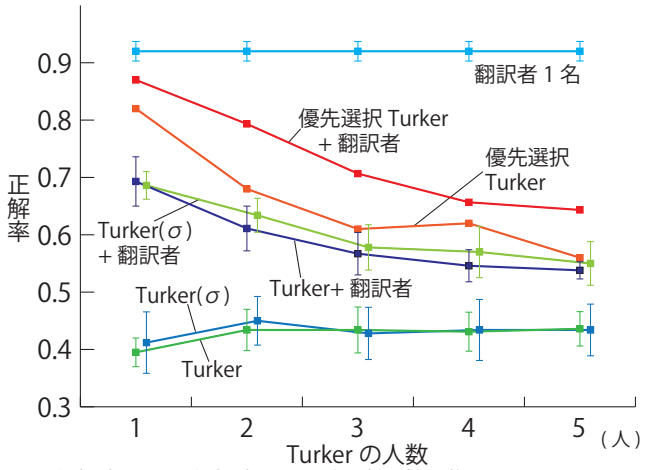
5.2　従来手法との比較

本節では，前節で述べた手法と従来手法との比較を行う．

Rzeszotarski らは我々と同様に，正解データを用いずに Turker の行動を利用した正確性評価を行っている [26]．文献 [26] では，Turker の作業時間や作業内容から Turker の質を判定している．また，Harrison らは作業時間が極端に短い Turker を除去している [27]．このように，Turker 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 [26] の手法では MTurk 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．本稿の提案手法では，MTurk から提供される情報のみで判定を行っている．また，人によって作業に必要となる時間は異なる．文献 [27] では単純な作業時間のみを考慮しているが，提案手法は入力の有無を用いて，Turker ごとに異なる，作業に必要な時間の影響を除去している点が異なっている．本節では，極端に長短である作業

*7　表 3 より，Turker1〜4 はそれぞれ 48 件，61 件，6 件，31 件のみ評価している．このため，4 人のデータを合わせた場合でも 7 件のテキストペアの評価を行うことができなかった．

*8　20 件以下の場合は極端に正解率が高くなるため．



※ 優先選択 Turker，優先選択 Turker+ 翻訳者の試行回数は 1 回である．

図 4　正確性評価の分析結果 (従来手法との比較)

Fig. 4　Analysis result of correctness evaluations (comparison of existing method and proposal method).

時間の結果を除いた手法 (従来手法) と提案手法との比較を行う．

本比較では，作業時間の標準偏差を閾値とし，外れ値[*9]となったものを除去したデータを利用した．その後，4.1 節の分析手法を適用したものを従来手法とし，比較を行った．比較結果を**図 4** に示す．図 4 中の「Turker(σ)」と「Turker+翻訳者 (σ)」が従来手法に当たる．なお，従来手法では 911 件のデータを用いている．また，「Turker」と「Turker+翻訳者」は 997 件のデータを用いた (4.2 節で述べたデータの除去を行っていない) 手法である．図 4 より，標準偏差を閾値とした手法では十分にデータの除去の効果が出ていないことが分かる．また，提案手法は従来手法よりも正解率が高く，良質な評価者を抽出できていることが分かる．

なお，文献 [26] では作業時間が正確に得られない場合が指摘されているが，本手法では入力の有無を用いて作業時間を用いているため，影響を減らすことができていると考えられる．

6.　おわりに

本稿では，正確な多言語間コミュニケーションのための用例対訳の正確性評価手法について検討を行った．本手法では，クラウドソーシングの作業者である Turker に正確性評価を依頼し，そのデータをもとに正確性評価手法の検討を行った．本研究の貢献は以下の 2 点である．

(1) タスク作業時間を元にした Turker の優先順位付けを行うことにより，翻訳者に準ずる正確性評価を可能にする手法を提案した．

(2) 一部の多言語テキストペアにおいては，正確性の高い評価を行う翻訳者と同等もしくは高い精度で Turker が評価を行えていることを示した．

[*9] (平均+標準偏差) を上限値，(平均+標準偏差) を下限値としたものを有効な作業時間とし，それ以外の値を外れ値とする．

今後は他のデータセットや Turker を用いて同様の結果が得られるか追加実験を行う．また，タスク作業時間以外の指標を用いた Turker の選択手法の検討を行う．

謝辞　本研究の一部は，科研費基盤研究 (B)(22300044) の助成を受けたものである．

参考文献

[1] 法務省：平成２２年末現在における外国人登録者統計について，法務省 (オンライン)，入手先〈http://www.moj.go.jp/nyuukokukanri/kouhou/nyuukantourokusyatoukei110603.html〉(参照 2012-05-17).

[2] 独立行政法人日本学生支援機構：平成 23 年度外国人留学生在籍状況調査結果，独立行政法人日本学生支援機構 (オンライン)，入手先〈http://www.jasso.go.jp/statistics/intl_student/data11.html〉(参照 2012-05-17).

[3] 法務省：平成２３年における外国人入国者数及び日本人出国者数について（確定値），法務省 (オンライン)，入手先〈http://www.moj.go.jp/nyuukokukanri/kouhou/nyuukokukanri04_00017.html〉(参照 2012-05-17).

[4] 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 2012-05-17).

[5] 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).

[6] 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).

[7] Kim, K. J. and Bonk, C. J.: Cross-Cultural Comparisons of Online Collaboration, *Journal of Computer Mediated Communication*, Vol. 8, No. 1 (2002).

[8] 高嶋愛里：在日外国人支援活動：京都における「医療通訳システムモデル事業」，国際保健支援会 2 (2005).

[9] 犬飼　章：第２回多文化共生の推進に関する意見交換会 (宮城県の取り組み事例)，総務省 (オンライン)，入手先〈http://www.soumu.go.jp/main_sosiki/kenkyu/tabunka/21171_3.html〉(参照 2012-05-17).

[10] 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).

[11] 杉田奈未穂，丸田洋輔，長谷川旭，長谷川聡，宮尾　克：ケータイ多言語対話システムとその応用，シンポジウム「モバイル'09」，pp. 63–66 (2009).

[12] 福島　拓，吉野　孝，重野亜久里：用例対訳を用いた多言語問診票作成システムの開発と評価，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. 2011-GN-78, No. 14, pp. 1–7 (2011).

[13] 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報告，人工知能と知識処理研究会，Vol. AI2010-47, pp. 19–24 (2011).

[14] 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，情報処理学会研究報告，コンシューマ・デバイス＆システム研究会，Vol. 2012-CDS-4, No. 5, pp. 1–8 (2012).

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

[16] Doan, A., Ramakrishnan, R. and Halevy, A. Y.: Crowdsourcing systems on the World-Wide Web, Vol. 54, No. 4,

pp. 86–96 (2011).
[17] Callison-Burch, C.: Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing(EMNLP2009)*, pp. 286–295 (2009).
[18] Matsuda, M. and Kitamura, Y.: Development of Machine Translation System for Japanese Children, *Proceedings of the 2009 ACM International Workshop on Intercultural Collaboration (IWIC'09)*, pp. 269–271 (2009).
[19] 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).
[20] 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，信学論，Vol. J88-D1, No. 9, pp. 1459–1466 (2005).
[21] 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).
[22] 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 (2012).
[23] Chen, D. L. and Dolan, W. B.: Collecting Highly Parallel Data for Paraphrase Evaluation, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 190–200 (2011).
[24] Negri, M. and Mehdad, Y.: Creating a Bi-lingual Entailment Corpus through Translations with Mechanical Turk: $100 for a 10-day Rush, *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 212–216 (2010).
[25] Walker, K., Bamba, M., Miller, D., Ma, X., Cieri, C. and Doddington, G.: Multiple-Translation Arabic (MTA) Part 1, *Linguistic Data Consortium, Philadelphia* (2003).
[26] Rzeszotarski, J. M. and Kittur, A.: Instrumenting the crowd: using implicit behavioral measures to predict task performance, *Proceedings of the 24th annual ACM symposium on User interface software and technology (UIST 2011)*, pp. 13–22 (2011).
[27] Harrison, C., Horstman, J., Hsieh, G. and Hudson, S. E.: Unlocking the Expressivity of Point Lights, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2012)*, pp. 1683–1692 (2012).