J-042

クラウドソーシングを用いた多言語用例対訳の正確性評価手法の検討
Proposal of Multilingual Parallel Texts Evaluation Method Using Crowdsourcing

福島　拓 †　　吉野　孝 †
Taku Fukushima　　Takashi Yoshino

1. はじめに

現在，世界的なグローバル化により，日本国内でも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，在日外国人や訪日外国人の中には，日本語を理解できない人が多数存在している [1]．一般に，母語以外の言語によるコミュニケーションは困難である [2]．このため，正確なコミュニケーションが求められる分野の支援には，用例を正確に多言語へと翻訳した多言語コーパスである「用例対訳」が利用されている [3]．我々は正確性が求められる医療分野の用例対訳の収集，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 [4]．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を用いた用例対訳の正確性評価を検討する．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 [5]，クラウドソーシングの利用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．しかし，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と考えられる．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．

また，本稿で提案する手法では，非専門家による評価者の評価の他に，翻訳者による評価を 1 名以上入れることとする．これは，実際の利用現場に用例対訳の提供を行う際に，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に行う．

2. クラウドソーシングを用いた正確性評価

2.1 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 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．なお，TackPad では，(i) 医療従事者や患者などの利用者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) 医療従事者や患者，翻訳者などが登録された用例や用例対訳の正確性評価を相互に行う，の手順で正確な用例対訳の作成を行っている．

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 [6] を利用して 5 段階評価を行った[1]．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，1 対の評価につき 15 円を翻訳者に支払っている[3]．

2.2 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 2.1 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．評価基準は前節と同様に Walker らの適合性評価を用いた．また，評価が 3 以下の場合，日本語を正しい英語に翻訳するよう依頼した．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]．

3. 分析と考察

3.1 正確性評価の分析方法

本節では，2.2 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 の人数と翻訳者の評価の重み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 1 名のデータを取得し，その評価を w 倍にする（重みをつける）．
(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，2.1 節のデータセットの正確・不正確と一

†和歌山大学

[1]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．
[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．
[4]https://www.mturk.com/
[5]約 4 円 (2012/7/2 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．

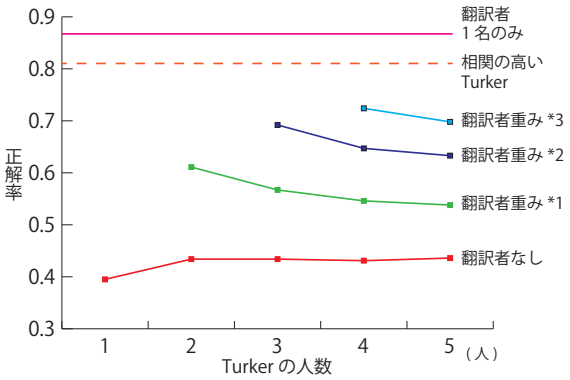

図 1: 正確性評価の分析結果

表 1: データセットと Turker の相関係数

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．

致しているかを調べる．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．

(4) (1)～(3) の操作を各対に対して行い，Turker が n 人，翻訳者の重みが w 倍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 1～5 名，評価者の重み w は 1～3 倍を用いた．

3.2 分析結果と考察

分析結果を図 1 に示す．なお，分析結果のうち，重み付けにより翻訳者が評価の半分以上を占める場合は除いている．図 1 より，翻訳者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．また，翻訳者を入れた場合，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．ただし，翻訳者 1 名の正解率の平均は 0.87 であった．

また，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表 1 に示す．表 1 より，データセットの評価との相関が低い Turker が多数存在していたことが分かる．このため，相関係数が低い Turker のデータを除去する必要があると考えられる．ただし，相関係数が高い Turker も存在していた．各データの評価のうち，最も高い相関係数であった Turker の評価を選択できたと仮定した場合，正解率は 0.81 だった (図 1 中の破線)．このため，高い相関係数を持つ Turker の評価と翻訳者の評価とを合わせることで，適切な正確性評価が可能であると考えられる．

次に，不正確なデータに関しての考察を行う．2.1 節のデータセットのうち，不正確なデータ (20 対) のみに限定した結果を図 2 に示す．また，翻訳者 1 名の正解率の平均は 0.83 であった．このことより，間違いデータに関しては比較的正確に判定できていることが分かる．

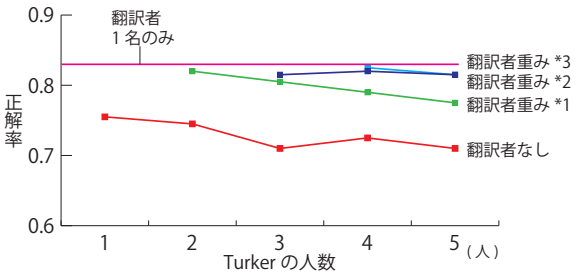

図 2: 正確性評価の分析結果 (不正確データのみ)

また，本稿で利用したデータセットにはスペルミスが含まれた文が 2 文存在していたが，人手 (翻訳者，Turker) による正確性評価では両方とも検出できない傾向にあった．データセットのスペルミスは，“子宮がん”–“uterrine cancer”[7]と，“肺炎”–“neumonia”[8]であった．“子宮がん”は Turker の 1 名のみが評価 2 をつけ，正しい単語を記述していた．しかし，翻訳者の 3 名と Turker の 9 名は 4 以上を付与しており，正確に判定できていなかった．また，“肺炎”は Turker の 3 名，翻訳者の 1 名が 3 以下の評価をつけたが，その他の評価者は 4 以上を付与していた．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．ただし，発見率が低いため，スペルミスの検出は機械的に行う必要があると考えられる．

4. 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利用者である 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，正確性の高い Turker の評価は専門家の評価と同等の精度であることを示したことである．また，多言語間の正確性評価手法の精度向上が可能であることを示したことである．

今後は Turker が正しく評価していないデータの除去作業を行い，正確性の向上を目指す．

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] 田村太郎：多民族共生社会ニッポンとボランティア活動，明石書店 (2000).

[2] Takano, Y., et al: A temporary decline of thinking ability during foreign language processing, Journal of Cross-Cultural Psychology, 24, pp.445-462(1993).

[3] 宮部真衣ほか：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，信学論，Vol.J92-D,No.6，pp.708-718(2009).

[4] 福島拓ほか：正確な情報共有のための多言語用例対訳共有システム，情処，2012-CDS-4(5)，pp.1-8(2012).

[5] Callison-Burch, C.: Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk, EMNLP 2009, pp.286-295(2009).

[6] Walker, K., et al: Multiple-Translation Arabic (MTA) Part 1, Linguistic Data Consortium, Philadelphia (2003).

[7]正しくは “uterine cancer”
[8]正しくは “pneumonia”