# クラウドソーシングによる翻訳評価の分析

後藤　真介[†a)]　　林　冬惠[†b)]　　石田　亨[†c)]

# Analysis of Translation Evaluation by Crowdsourcing

Shinsuke GOTO[†a)], Donghui LIN[†b)], and Toru ISHIDA[†c)]

**あらまし**　近年，機械翻訳サービスの増加により翻訳評価の需要が高まっている．しかし，これまで一般的に行われてきたバイリンガルの専門家による翻訳評価は時間的及び金銭的なコストが高くかかるという問題がある．本研究ではクラウドソーシングを用いた翻訳評価を提案する．クラウドソーシングは Web 上の不特定多数の作業者にタスクを依頼することで低価格かつ高速に作業を実行できるが，翻訳評価タスクは専門性を要するタスクであり，作業者にバイリンガル能力を要する翻訳タスクは実行されていない．そこで，クラウドソーシングを用いた翻訳評価に関する分析を行うために専門家の翻訳評価の用途に基づく三つの比較指標を定義し，各指標について専門家とクラウドソーシング作業者の評価を比較した．また，NTCIR-9 PATENT データセットを用いた専門家と同じ手法による中英翻訳評価タスクの設計を行い，実際のクラウドソーシングプラットホームである Amazon Mechanical Turk (AMT) において実験を行った．実験の結果，複数の原文からなる文書集合に対する機械翻訳システムの相対評価に関してクラウドソーシングが有用であることが示された．

**キーワード**　クラウドソーシング，翻訳，評価

## 1. まえがき

近年，多くの機械翻訳システムが実用的に用いられるようになってきている．しかし，機械翻訳は人手による翻訳に比べて翻訳品質が保証されない問題があり，翻訳文の評価が必要となる．これまで，翻訳評価は主にバイリンガルの翻訳専門家によって行われてきた．専門家による評価指標は翻訳の意味を表す正確さと，翻訳文が翻訳先の言語においてどれくらい文法的に正しいかを表す流暢さの二つが存在する [1]．しかし，このような専門家による評価は高いコストを要する．

機械翻訳に対する翻訳評価の目的には各翻訳文の評価値の取得，ある原文に対するそれぞれの翻訳文の相対評価，複数の原文からなる文書集合に対する機械翻訳システムの相対評価などが存在する [2]〜[4]．それぞれの評価の用途は順番に専門家による各翻訳に対するスコアの利用，ある文に対する最良の翻訳結果の選択，そして複数の入力からなる最良の機械翻訳システムの選択である．一方で，翻訳の自動評価を目的とした研究も複数存在する．例えば BLEU や METEOR などは，機械翻訳文と正解となる参照訳との類似性に基づいた翻訳の評価を行っている [2], [5]．これらの評価値は専門家の評価と比較した場合に高い相関係数が得られているが，翻訳文ごとの品質の絶対値の評価を行うことは困難である．また，これらの翻訳評価のためには一般的に複数の参照訳を必要とする．原文及び参照訳が事前に与えられない場合，これらの自動評価手法を適用させることは難しい．

そこで，これまでの専門家や自動評価手法に代わってクラウドソーシングを用いた翻訳評価を提案する．クラウドソーシングとは Web 上の不特定多数の作業者に仕事を依頼する新たな形の雇用形態で，同じ作業を専門家が実行する場合に比べて低コストかつ迅速な作業が可能となる．既存のクラウドソーシングを用いた翻訳評価はタスク実行時に専門家の参照訳が用いられている．これによりモノリンガルによる作業を実現しているが，このようなタスクの作成には専門家による翻訳が必要である．

本研究は，クラウドソーシングによる翻訳評価が専門家の翻訳評価の各用途に対して，有用な評価が得られるかどうかの分析を目指す．非専門家による翻訳評

† 京都大学大学院情報学研究科，京都市
Graduate School of Informatics, Kyoto University, Yoshida Honmachi, Sakyo-ku, Kyoto-shi, 606–8501 Japan
a) E-mail: s-goto@ai.soc.i.kyoto-u.ac.jp
b) E-mail: lindh@i.kyoto-u.ac.jp
c) E-mail: ishida@i.kyoto-u.ac.jp

 

価が実現すれば，専門家を用いることなく翻訳の評価が可能であると同時に，参照訳の存在を前提としない翻訳評価は機械翻訳の評価だけでなく，近年広がっているクラウドソーシングをはじめとした非専門家による翻訳の評価にも活用可能である．この目的に向け，我々は専門家による翻訳評価の用途に応じた三種類の比較指標を定義し，クラウドソーシングプラットホームである Amazon Mechanical Turk (AMT)(注1) を通じてクラウドソーシングの作業者に専門家と同じ基準である正確さと流暢さの二つの指標による評価を依頼する．その後，定義された指標それぞれに対してクラウドソーシングの評価と専門家の評価を比較し，その結果を分析する．

## 2. 関連研究

これまで，翻訳評価は人手によって行われてきた．一般的に，翻訳評価値は正確さと流暢さの二つの指標で構成される．正確さは翻訳された文がどの程度原文の意味を正しく伝えているかを表し，一方で流暢さは翻訳された文がその言語の文としてどれくらいふさわしいかの指標である [1]．同時に，翻訳自動評価手法が複数提案されている．広く知られている自動評価手法の例は BLEU や METEOR が挙げられる [2], [5]．これらの自動評価は，複数の参照訳と機械翻訳文の間の N-グラムの一致度に基づく手法を用いている．一方で，自動翻訳評価手法の問題を指摘する論文も複数存在する．Ying らの研究でははは BLEU と NIST はそれぞれ専門家との評価において翻訳システム間のランキングが異なるという実験結果が得られている [6]．

また，クラウドソーシングは様々な評価のために用いられてきた．例えば Chen らは動画を配信する際の圧縮度と体験品質をクラウドソーシングの作業者に測定させた [7]．作業者は品質の異なる 6 種類の動画の主観的な評価を一対比較し，全ての一対比較による評価結果を正規化することで評価値として出力する．また，Gabriella らは書籍の検索結果の評価をクラウドソーシング作業者を通じて行っている [8]．このタスクでは，作業者は検索内容と書籍の一致度を複数の設計のタスクで評価させ，クラウドソーシングの品質管理をきちんとタスクに組み込むことが評価品質の向上につながると結論付けている．

クラウドソーシングを用いた翻訳に関する研究も存在する．Omar らはクラウドソーシングを用いた翻訳を実現するために複数の作業者に翻訳文を作成させ，その後複数の翻訳文候補から後編集を行い，そして最後に投票によって最適な翻訳を選択する枠組みの有用性を示した [9]．Vamshi らは統計的機械翻訳の品質向上のために能動学習とクラウドソーシングによる翻訳文作成を組み合わせた翻訳システムを提案した [10]．

クラウドソーシングを用いた翻訳評価に関する先行研究としては，Callison-Burch ら及び Luisa らが翻訳の評価をクラウドソーシングに依頼している [11], [12]．Callison-Burch らは専門家の作成した翻訳文を利用した評価を行い，クラウドソーシングの翻訳評価が有用であると結論付けている．Luisa らの研究ではクラウドソーシング実験において原文，参照訳，そして翻訳文を用意し，機械翻訳システム間のランキングを作成している．これらの研究は，専門家の作成した参照訳を前提とし，モノリンガルでも作業可能なタスクに対してクラウドソーシングが有用に働くことを示している．

本研究では参照訳が存在しない場合に非専門家である作業者の評価結果の分析を行う．クラウドソーシングの作業者には専門家と同じ指標の 5 段階評価を依頼し，またクラウドソーシングと専門家の評価結果を比較することで分析を行っている．

## 3. クラウドソーシングと専門家の比較指標

本研究では，専門家による翻訳評価の用途に応じた 3 種類の手法を用いてクラウドソーシングと専門家の評価を比較する．参照訳が存在しない場合のクラウドソーシングによる各評価の分析により，翻訳評価の利用方法の決定が可能である．

それぞれの評価値に対する比較手法を本研究では translation-score, sentence-score, system-score の比較と呼ぶ．translation-score は各翻訳文の絶対値による評価を意味し，この評価は機械翻訳及び翻訳評価手法の評価に用いられている [2]．また，sentence-score は原文に関する相対評価であり，Shi らは 300 文の原文に対して機械翻訳の評価値の順位による比較を実施している [3]．最後に，system-score は機械翻訳システムごとの評価値であり，Goto らは 300 文の原文それぞれに対する 23 個の機械翻訳文を評価し，その評価値の平均を翻訳システムの評価として分析を行っている [4]．我々は，ここで述べたそれぞれの評価値について専門家による評価とクラウドソーシングによる評

(注1)：http://www.mturk.com

価を比較する．

次に，本研究で比較する翻訳評価値の定義を行う．$m$ 個の原文とそれぞれの原文に対する $n$ 個の機械翻訳が存在すると仮定する．このとき，$sc_{ij}$ は $i$ 番目 $(1 \leq i \leq m)$ の原文を $j$ 番目 $(1 \leq j \leq n)$ の機械翻訳が出力した翻訳文のクラウドソーシングによる評価値となる．本研究では，複数人のクラウドソーシングの作業者が同じ翻訳文に対して評価を行うが，それぞれの翻訳文に対する評価の平均をクラウドソーシングによる評価値とみなす．同様に，$sp_{ij}$ は専門家による評価値である．詳細は後述するが，今回実験に用いたデータセットでは専門家による評価はそれぞれの翻訳文に対し一つ存在するため，その値を専門家の評価とする．

これらの定義に基づき，translation-score, sentence-score, system-score のそれぞれを以下の式で定める．以下で $Sc$ はクラウドソーシングによって得られたスコアのベクトルを表し，$Sp$ は専門家が評価したスコアのベクトルを表す．定義では $Sc$ のみを記述するが，$Sp$ も各評価の $sc$ を $sp$ に置き換えることにより，同様に定義される．

Translation-score：クラウドソーシングによる Translation-score を表すベクトルである $Sc_{translation}$ は式 (1) によって定義される．

$$Sc_{translation} = \{sc_{11}, sc_{12}, \ldots, sc_{1n}, \ldots, sc_{mn}\} \quad (1)$$

Translation-score の比較はクラウドソーシングと専門家の評価の絶対値の一致度を表す．この評価はそれぞれの翻訳スコアの平均絶対誤差 (MAE) を取得することによって行われる．MAE は二つのベクトルの各要素に対して誤差の絶対値をとり，その平均によって計算される．今回，クラウドソーシングと専門家の評価の比較として MAE$(Sc_{translation}, Sp_{translation})$ 計算する．

Sentence-score：Sentence-score $Sc_{sentence}(i)$ は原文 $i$ に各機械翻訳システムが翻訳した文に対するクラウドソーシングによる評価のベクトルであり，式 (2) で表現される：

$$Sc_{sentence}(i) = \{sc_{i1}, sc_{i2}, \ldots, sc_{in}\} \quad (2)$$

$Sc_{sentence}(i)$ と $Sp_{sentence}(i)$ の相関係数をとることにより，専門家とクラウドソーシングの間の原文 $i$ に関する相対評価の一致度が求められる．

System-score：クラウドソーシングによる System-score $Sc_{system}$ はクラウドソーシングがそれぞれの機械翻訳システムに対して付けた入力原文全てに対するスコアの平均のベクトルであり，式 (3) で求められる：

$$Sc_{system} = \{\frac{\sum_{i=1}^{m}(sc_{i1})}{m}, \ldots, \frac{\sum_{i=1}^{m}(sc_{in})}{m}\} \quad (3)$$

ベクトルの各要素は与えられた全ての原文に対するある機械翻訳の評価値の平均を表し，$Sc_{system}$ と $Sp_{system}$ の相関係数は与えられた全ての原文を文書集合とした場合の機械翻訳システムの相対評価の類似度を示す．

## 4. クラウドソーシングによる翻訳評価タスクの設計

**3.** で定義した比較手法によってクラウドソーシングの有用性を分析するため，我々は翻訳の評価値を実際に取得するための実験を行った．実験では中英翻訳評価を行い，クラウドソーシングによる翻訳の評価の実用性を評価した．本実験では AMT をクラウドソーシングプラットホームとして用いた．AMT は世界最大のクラウドソーシングプラットホームの一つである．また，また，実験に必要な中英翻訳データとして NTCIR-9 PATENT データセットを利用した．NTCIR-9 PATENT データセットは日英中の 3 言語の国際特許を翻訳するために作成されたデータセットである [4]．原文は国際特許から文ごとに抽出され，極度に類似した文が存在しないようになっている．各原文に対する 23 システムの機械翻訳結果を，それぞれの翻訳に対して一人のバイリンガル専門家が正確さと適切さによる 5 段階評価を行っている．事前にバイリンガル専門家は 3 人で 100 文の翻訳評価を行い，それぞれに対して共通の評価基準ができるように直接話し合って訓練を行っている．また，バイリンガルは参照訳を見ることなく，原文の中国語とそれに対する英語の機械翻訳の出力を見て評価値を決定している．加えて，英語を母語とする専門家によって流暢さの測定を行った．これらのデータセットに対し，本実験においてクラウドソーシングによる評価を行い，その有用性を検証する．実験では NTCIR-9 PATENT データセットから 10 文の原文を抽出し，それぞれに対する 23 翻訳システムによる翻訳文合計 230 文に対して評価を行う．

中英翻訳に対する品質評価タスクを行う理由は AMT

表 1　翻訳文評価タスクの例
Table 1　Example of translation evaluation task.

| 中国語 | 正確さ | 流暢さ |
|---|---|---|
| 障碍物的一个可行的实施方案为在鼓内的螺旋状障碍物(图1中未示出)。 | 5 (全ての意味が一致) | 5（完璧な英語） |
| | 4 (ほとんどの意味が一致) | 4（良い英語） |
| English | 3 (大体の意味が一致) | 3（非ネイティブの英語） |
| One possible embodiment is within the drum of an obstacle (not shown in Fig. 1). | 2 (意味がほとんど一致しない) | 2（流暢でない英語） |
| | 1 (全く一致しない) | 1（理解不能） |

の作業者数である．中英翻訳評価の作業者となりうる在米中国人は在米日本人に比べて圧倒的に数が多く，作業完了に必要な時間に大きな違いが発生する．予備実験で行った日英翻訳の評価実験では，5 日間でタスクを実行した作業者数は 6 名であったため，十分な作業者が得られないと判断した．

表 1 に実施した評価実験の具体例を示す．評価実験では AMT を用い，専門家と同じ指標によるクラウドソーシングによる評価を実行した．実際のタスクでは，一つの原文に対する 23 翻訳全ての評価を行うように依頼した．翻訳評価は各翻訳文に対する絶対評価だけでなく，同じ原文に対する翻訳間の相対的な指標としても扱われるため，作業者が全ての翻訳文を見ることにより一貫性を保っている．作業者には，原文と翻訳文を読み，その後正確さあるいは流暢さによる評価を行うように指示を行った．正確さ及び流暢さの定義は以下のとおりである．

正確さ：原文が表す意味をどの程度翻訳文が伝えているか

流暢さ：翻訳文が英語のネイティブ文と比較してどの程度正しい英語であるか

また，本実験ではスパマーを排除するために品質管理を行った．スパマーとは自身の利益のためにタスクに対して誠実な回答を行わない作業者やタスクの自動実行プログラムを意味する．

本研究では二つの手法でスパマーに対処した．一つはアクセス可能な地域の制限であり，もう一つは資格テストである．

前者に関しては，翻訳評価タスクを実行可能な作業者はアメリカからのアクセスに制限した．これは，タスクを行う上では英語と中国語の両方を知っている必要があるが，AMT の作業者は中国からアクセスすることはほとんどないためである．Joel らは，AMT の作業者の 92%はアメリカあるいはインドに集中していると説明している [13]．もう一つの理由として，予備実験ではアメリカ以外からの作業者が品質が低い作業結果を提出したことが挙げられる．上記の理由のため，作業者のアクセス地域を制限することは品質向上につながると判断した．

もう一つの対処である資格テストは，作業を希望するクラウドソーシング作業者に対して課する資格テストである．もし作業者が資格テストに対して十分な点数が得られなかった場合，タスクを実行することはできない．資格テストは本番のタスクの評価と同じ 5 段階評価によって行われる．原文と翻訳文として，NTCIR-9 データセットから，実際に依頼するタスクと異なる原文 1 文とそれに対する翻訳文 10 文を抽出した．作業者は 10 個の翻訳に対して 5 段階評価を行い，正解である専門家の評価との相関係数によって作業者の資格の有無を判断する．正確さ及び流暢さに関する資格テストはそれぞれ独立に実施され，異なる作業者が雇われた．本実験では相関係数が 0.4 以下の作業者を排除することで作業者の能力を担保した．相関係数の基準として 0.4 を設定した理由は，クラウドソーシングの作業者として適切に作業を行うことが可能であるという判断のためである．作業者がスパマーではなく，かつ一定の翻訳に対する知識を保証することが今回のタスク実行のためには必要である．複数の能力をもつクラウドソーシング作業者は十分な数が存在しないため，作業者の資格として翻訳評価に関する能力のみを判定している．

正確さ評価と流暢さ評価のそれぞれの実験に対し，14 人と 10 人が参加し，翻訳評価を行った．これらによって，計 4,347 個の翻訳に対する評価が行われた．また，一人当たり平均 7 個のタスクを実行し，5 人の作業者が 10 文全ての原文に対して正確さあるいは流暢さに対して評価を行った．

## 5. クラウドソーシングによる翻訳評価の分析

本章では **4.** で述べた実験の結果及びクラウドソーシングによる評価結果の分析について説明する．

まず，translation-score の比較結果は正確さ及び流暢さにおいて，それぞれ MAE が 0.63，0.68 となっ

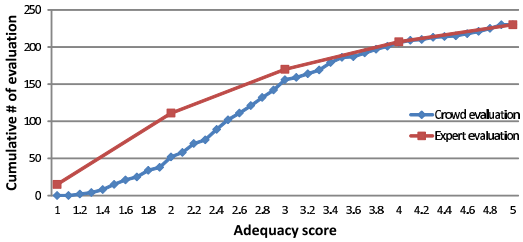


図 1　専門家とクラウドソーシングの正確さ評価値の累積グラフ

Fig. 1　Cumulative graph of professional and crowdsourcing evaluation of adequacy.

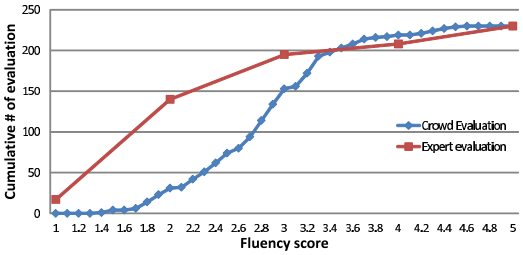


図 2　専門家とクラウドソーシングの流暢さ評価値の累積グラフ

Fig. 2　Cumulative graph of professional and crowdsourcing evaluation of fluency.

た．また，絶対誤差が 0.5 より小さく，四捨五入すると専門家と同じ評価値になる評価結果は正確さ及び流暢さでそれぞれ 106/230(46%)，98/230(43%) であった．このため，クラウドソーシングによる評価値を単純に専門家の評価として当てはめるだけでは専門家の評価を代替できない．図 1，図 2 は translation-score の専門家のスコアとクラウドソーシングのスコアの比較の累積グラフを示している．この図からは，正確さ，流暢さ共に専門家の評価値が 2 であるときにクラウドソーシングの評価値は 3 に近いものになっていることが多い．このことから，作業者は専門家の評価が低い場合に専門家に比べてやや高い評価を行っていることが分かる．一方で，専門家の評価が高い場合，正確さの場合は専門家の評価値が 4 や 5 のときは累積度数が同程度になっている．これに対し流暢さでは専門家が 4 や 5 の評価を付ける場合にやや低い評価を行っている．

次に，sentence-score の比較はそれぞれの原文に対する順位付けの有効性を検証する．図 3 から，sentence-score の相関係数は原文及び評価指標ごとに大きな差があることが分かる．正確さについて平均の相関係数は 0.62 であるが，最大で 0.92（原文 ID8），最小で

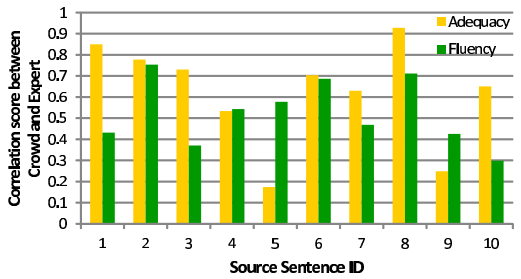


図 3　それぞれの原文に対する専門家とクラウドソーシングによる評価の相関係数の棒グラフ

Fig. 3　Bar graph of correlation between expert and crowdsourcing for each source sentence.

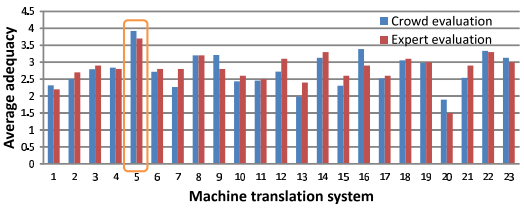


図 4　System-score それぞれの正確さ評価値の棒グラフ．クラウドソーシングによる評価は専門家による最良の翻訳システムである機械翻訳システム 5 を選択している

Fig. 4　Bar graph of adequacy for each translation system.　Crowdsourcing can select the best system, machine translation 5.

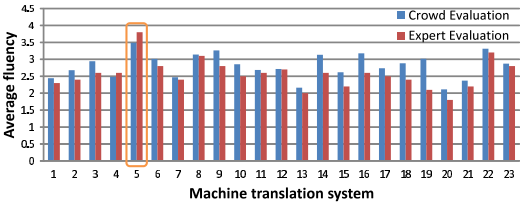


図 5　System-score それぞれの流暢さ評価値の棒グラフ．クラウドソーシングによる評価は専門家による最良の翻訳システムである機械翻訳システム 5 を選択している

Fig. 5　Bar graph of fluency for each translation system.　Crowdsourcing can select the best system, machine translation 5.

0.17（原文 ID5）と大きな差がついている．流暢さについても同様の傾向があるが，平均の相関係数が 0.53，最大値が 0.75（原文 ID2），最小値 0.30（原文 ID10）となり，正確さに比べて差は小さくなっている．これらの相関係数の大きな差の原因として，個々の原文に対する評価値のずれが相関係数に大きな影響を与えたことが考えられる．

また，原文 ID5 の正確さ評価を行った作業者からは

原文の意味を読み取ることが困難であったとの意見もアンケート結果から得られ，今後の課題として翻訳文だけでなく原文に対する評価も考慮する必要がある．

最後に，クラウドソーシングと専門家それぞれのsystem-scoreの比較では，相関係数が0.82，0.80となり，評価の相関が強いことが明らかになった．また，クラウドソーシングの評価の各原文に対する平均は，正確さ，流暢さ共に23個の機械翻訳システムから最適なものを選択することに成功している．図4及び図5ではクラウドソーシングによる評価が専門家によって最も翻訳品質が良いと判断された機械翻訳システム5を選択することに成功していることが分かる．この結果からはクラウドソーシングが複数の原文に対する最適な翻訳システムを選択することが可能であり，また複数の原文に対する相対的な評価を行う用途では利用可能なことが分かる．

## 6. 議　　論

クラウドソーシングによる翻訳評価実験の結果と専門家による評価を比較した結果，system-scoreの有用性が明らかになった一方で，sentence-score及びtranslation-scoreについては直接の利用は困難であるという結果が得られた．本章では具体的な翻訳文から評価の分析を行うとともに，クラウドソーシングの実用化の上で重要な金銭的コスト及び時間的コストについても言及する．

クラウドソーシングによる評価が専門家の評価と異なっている具体的な例を表2に挙げる．これらの異なった翻訳は専門家とクラウドソーシングの評価の仕方の違いによって引き起こされると考えられる．表2は専門家とクラウドソーシングの評価が大きく異なっている翻訳二つを抽出したものである．上の翻訳では，参照訳にある “schematized” や “package” の意味が失われているために専門家の評価値は2と低くなっている．しかし，翻訳文には大きな欠落がないため，作業者の評価値の平均は3.5となっている．一方，下の翻訳は，翻訳文中で太字で示してある “instantiated schematized” や “Examples” が文法的に誤っている．しかし，意味は全て訳出されているために専門家による評価値は5となっている．これは，機械翻訳の正確さは原文の伝えている内容を翻訳文が伝えている割合で定義され，翻訳文の文章としての正しさである流暢さとは異なる評価が行われるためである．一方で，クラウドソーシングの作業者は翻訳の正確さ評価についても文法を過大評価するために平均の評価値が2.83と悪くなっている．流暢さについて，表3では原文

表2　正確さ評価における専門家とクラウドソーシングの評価が異なる例
Table 2　Example sentences of different evaluation between professional and crowdsourcing by adequacy.

| 原文 | 機械翻訳 | 専門家正確さ | クラウド正確さ |
|---|---|---|---|
| 可以理解，上述系统300确保无论何时客户端应用程序实例化模式化类型的实例以便保存在存储中，该客户端就能访问模式包。 | It is to be understood that the above-described system 300 ensures that whenever a client application instanciation modes, examples of the types to be stored in the memory, the client can access mode. | 2 | 3.5 |
| Reference: It is to be understood that the above described system 300 ensures that whenever the client application instantiates an instance of a schematized type for persistence in a store, the client has access to the schema package. | It is to be understood that the above-described system 300 to ensure that whenever the client application **instantiated schematized** types of **Examples** to stored in storage , the client can access mode packets . | 5 | 2.83 |

表3　流暢さ評価における専門家とクラウドソーシングの評価が異なる例
Table 3　Example sentences of different evaluation between professional and crowdsourcing by fluency.

| 原文 | 機械翻訳 | 専門家流暢さ | クラウド流暢さ |
|---|---|---|---|
| 因为X3和X4不穿过信号变换单元，所以各自的声道配置信息为0。 | Since X3 and X4 does not pass through the signal conversion unit, so that respective channel configuration information 0. | 1 | 3.22 |
| Reference: Since X3 and X4 do not pass through signal converting units, each channel configuration information becomes 0. | Because X3 and X4 do not pass through the signal translation unit, therefore the respective sound track configuration information is 0. | 4 | 2.6 |

ID6 の文章の翻訳において，後半の節に動詞がない翻訳文になっているために専門家の評価が「理解不能」となっている．これに対し，前半の節は文章として大きな間違いはなく，クラウドソーシング作業者は前半の文としての正しさを考慮しために流暢さの平均値が 3.22 となっている．また，表下段の翻訳は専門家によって「良い英語」である 4 の評価を得ているが，作業者は “sound track” の翻訳の正確さが間違っていることを考慮したために流暢さ評価が 2.6 になっている．このことから，正確さと流暢さそれぞれの評価指標において，作業者が混同した評価を行わないような指示が必要となる．また，これらの評価値の違いには，特許文書という文書の領域に対する知識が少ないことが理由となって発生する場合も存在する．今後の改善として，特許文書に対する知識をもつモノリンガル作業者と，言語に対する知識をもつバイリンガル作業者を連携させ，専門的用語の訳出と翻訳品質をそれぞれ独立して検証する手法が考えられる．

また，クラウドソーシングによる翻訳評価を実用的に利用するためには，専門家と比較した場合の評価の品質だけでなく，金銭的費用やかかる時間についての議論も必要である．

金銭的費用については，本実験では一人当たり 23 翻訳の評価に 2 ドルを費やしている．これを作業者の時給に換算すると約 18 ドルとなった．作業者に専門的な能力を要しないアンケートタスクの時給が 1.71 ドルであることから，クラウドソーシングではこの報酬は高額である [14]．これは今回の実験では金額の議論をせずにより多くの作業者を求めたためである．また，専門家との比較では，日本翻訳者協会(注2)において英日翻訳に関して日本語 400 文字当たり 2000 円以上の相場であるという記述がある．このため，クラウドソーシングの作業者へ支払う費用は専門家に対する費用より安くなると考えられる．

時間的コストについては，今回は難易度の高い資格テストを実施したために 2 週間で 14 人の作業者しか集まらない結果となっている．また，この値は様々な要因に影響を受ける．一例をあげるだけでも，金銭的費用，翻訳言語対，タスクの内容，資格テストの難易度などである．特に **4.** で述べたとおり，AMT の作業者の大部分はアメリカまたはインドからのアクセスであるので，原文及び翻訳文の両者が英語以外の言語である場合，必要な作業者が集まることは考えにくい．この問題については，別のクラウドソーシングプラットホームを通じて評価を依頼するなど，他と異なるアプローチによる解決策が必要となる．

## 7. むすび

これまで多くの翻訳評価手法が提案されてきた一方で，既存の手法はコストや評価品質の面において問題を抱えていた．本研究ではクラウドソーシングを翻訳評価に適用させ，複数の評価指標によってクラウドソーシングによる評価が有効に働く用途を調査した．その結果として，system-score の比較から，複数の原文の評価値の平均値を用いて機械翻訳システムを比較する場合は，クラウドソーシングの評価値が専門家の評価に比べて相関係数が高く，利用可能であることが示された．その一方で translation-score や sentence-score の結果から，全ての翻訳文に対する評価が専門家と一致するとは限らないことが明らかになった．すなわち，専門家による参照訳の存在しない場合のクラウドソーシングによる翻訳評価は翻訳文が増加するほど専門家の評価に近づくことを示唆している．この結果を利用することにより，クラウドソーシングや機械翻訳に対するフィードバックとして活用することが期待される．

今後の課題として，今回の分析から得られた translation-score や sentence-score の問題を解決するために，クラウドソーシングの作業者それぞれの分析を詳細に行い，より専門家の評価とクラウドの評価を一致させる手法を提案する点が挙げられる．これを実現するためには資格テストの結果の利用や，機械学習に基づくフィッティング手法などが必要となるだろう．また本論文では言及していないが，作業者が実行する翻訳評価の一つのタスクあたりの分量を減らすことによる影響の分析が未実施であり，クラウドソーシングによるタスク実行の更なる効率化のために考えていかなければならない．

**謝辞** 本研究は，日本学術振興会科学研究費基盤研究 (S) (24220002，平成 24 年度～28 年度) 及び科学技術振興機構「問題解決型サービス科学研究開発プログラム」の補助を受けた．

## 文　献

[1] J. White, T. O’Connell, and F. O’Mara, “The arpa mt evaluation methodologies: Evolution, lessons, and future approaches,” Proc. 1994 Conference, Association for Machine Translation in the Americas,

(注2)：http://jat.org/ja/working_with_translators/

pp.193–205, 1994.
[2] S. Banerjee and A. Lavie, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp.65–72, 2005.
[3] C. Shi, D. Lin, M. Shimada, and T. Ishida, “Two phase evaluation for selecting machine translation services,” Proc. Eighth International Conference on Language Resources and Evaluation (LREC’12), pp.1771–1778, 2012.
[4] I. Goto, B. Lu, K.P. Chow, E. Sumita, and B.K. Tsou, “Overview of the patent machine translation task at the ntcir-9 workshop,” Proc. NTCIR, vol.9, pp.559–578, 2011.
[5] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu, “Bleu: A method for automatic evaluation of machine translation,” Proc. 40th annual meeting on association for computational linguistics, pp.311–318, Association for Computational Linguistics, 2002.
[6] Y. Zhang, S. Vogel, and A. Waibel, “Interpreting bleu/nist scores: How much improvement do we need to have a better system?,” Proc. Fourth International Conference on Language Resources and Evaluation (LREC 2004), pp.2051–2054, 2004.
[7] K.T. Chen, C.J. Chang, C.C. Wu, Y.C. Chang, and C.L. Lei, “Quadrant of euphoria: A crowdsourcing platform for qoe assessment,” IEEE Netw., vol.24, no.2, pp.28–35, 2010.
[8] G. Kazai, J. Kamps, M. Koolen, and N. Milic-Frayling, “Crowdsourcing for book search evaluation: Impact of hit design on comparative system ranking,” Proc. 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp.205–214, 2011.
[9] O. Zaidan and C. Callison-Burch, “Crowdsourcing translation: Professional quality from non-professionals,” ACL, pp.1220–1229, 2011.
[10] V. Ambati, S. Vogel, and J.G. Carbonell, “Active learning and crowd-sourcing for machine translation,” LREC, vol.11, pp.2169–2174, Citeseer, 2010.
[11] C. Callison-Burch, “Fast, cheap, and creative: Evaluating translation quality using amazon’s mechanical turk,” Proc. 2009 Conference on Empirical Methods in Natural Language Processing, vol.1, pp.286–295, Association for Computational Linguistics, 2009.
[12] L. Bentivogli, M. Federico, G. Moretti, and M. Paul, “Getting expert quality from the crowd for machine translation evaluation,” Proc. MT Summmit, vol.13, pp.521–528, 2011.
[13] J. Ross, L. Irani, M. Silberman, A. Zaldivar, and B. Tomlinson, “Who are the crowdworkers?: Shifting demographics in mechanical turk,” CHI’10 Extended Abstracts on Human Factors in Computing Systems, pp.2863–2872, 2010.
[14] G. Paolacci, J. Chandler, and P.G. Ipeirotis, “Running experiments on amazon mechanical turk,” Judgment and Decision making, vol.5, no.5, pp.411–419, 2010.

（平成 25 年 10 月 30 日受付，26 年 2 月 24 日再受付）

**後藤　真介**

2013 京都大学大学院情報学研究科社会情報学専攻修士課程了．現在，同大学院社会情報学専攻博士課程在学中．人工知能，クラウドソーシングに興味をもつ．

**林　　冬惠**　（正員）

2005 中国上海交通大学計算機科学専攻修士課程了．2008 京都大学大学院情報学研究科社会情報学専攻博士課程了．博士（情報学）．現在，同大学社会情報学専攻特定助教．サービスコンピューティングに興味をもつ．

**石田　　亨**　（正員：フェロー）

1976 京大・工・情報工学卒，1978 同大大学院修士課程了．同年日本電信電話公社電気通信研究所入所，この間，ミュンヘン工科大学，パリ第六大学，メリーランド大学，上海交通大学，清華大学客員教授などを経験．工博．情報処理学会，IEEE フェロー．現在，京都大学大学院情報学研究科社会情報学専攻教授．デジタルシティ，言語グリッド，異文化コラボレーションなど情報技術と社会をつなぐ研究プロジェクトを推進．