クラウドソーシング上の単言語話者に依頼可能な多言語用例対訳作成手法の提案と評価

福島 拓
和歌山大学大学院システム工学研究科
fukushima@yoslab.net

吉野 孝
和歌山大学システム工学部
yoshino@sys.wakayama-u.ac.jp

1　はじめに

我々は用例対訳[1]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 [9]．本システムでは用例対訳の収集数が十分でないという問題を抱えている．本システムでは多言語対訳の作成を翻訳者が行っているが，翻訳者の人数は少なく，大きな負担がかかっている．

そこで本稿では，単言語話者のみで多言語用例対訳候補の作成可能とする手法を提案する．その際，画像を媒体とすることで，正確な対訳作成を目指す．また，クラウドソーシング [2, 3] 上で単言語話者へ作業委託を行う．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利点がある．しかし，クラウドソーシング上で多言語が関係する作業委託を行った場合，特に不正確なものが多く含まれることが分かっている [1, 5, 7]．このため，本稿では多言語による悪影響を減らすために，クラウドソーシングへの作業委託を単言語で行うこととする．

2　関連研究

クラウドソーシングを用いた多言語データの収集はいくつか行われている．Callison-Burch はクラウドソーシングを用いた多言語対の正確性評価を [1]，Negri らはクラウドソーシングを用いた多言語対の作成 [5] をそれぞれ行っている．これらの研究では，多言語話者を対象としており，両言語の文を見せた正確性評価や，一方の言語を提示してもう一方の言語への翻訳の依頼をそれぞれ行っている．また，不適切に対価を得ようとするクラウドソーシング上の労働者が存在することを考慮した手法がそれぞれ提案されている．また，我々も翻訳者の作業特徴を考慮し，作業時間をもとにした多言語対の正確性評価手法を提案した [7]．

しかし，これらの研究では，クラウドソーシング上に少ないと考えられる多言語話者を対象とした作業委託を行っている．2 言語が関係するこれらの作業は比較的難解な作業となる．このため，単純な作業で対価を得ようとする労働者が多いと考えられるクラウドソーシング上では，不適切に対価を得ようとする労働者が多く作業を行う可能性が考えられる．

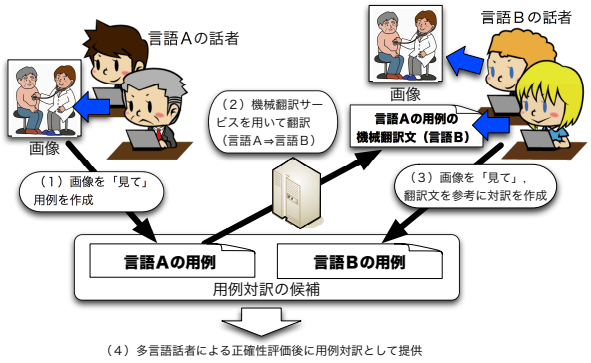

図 1: 提案手法の流れ

そこで我々は，多言語の作業タスクを単言語に分割して作業委託を行うことで，対象となる労働者数を増やすこととした．さらに，従来よりも単純な作業とすることで，不適切な作業者を減らすことを目指す．

3　提案手法

図 1 に提案手法の流れを示す．本手法は下記の 4 ステップで構成されている．

Step 1 用例の作成
図 1(1) で，言語 A を母語とする利用者が画像をもとに言語 A の用例の作成を行う．

Step 2 翻訳
図 1(2) で，言語 A の用例を機械翻訳サービスで言語 B へ翻訳を行う．

Step 3 対訳の作成
図 1(3) で，言語 B を母語とする利用者が翻訳文の正確性判定と，機械翻訳文をもとに正しい言語 B の用例作成（手法上は対訳作成）を行う．この時，図 1(1) で言語 A 話者へ提示した画像も合わせて表示することで，適切性を高めた言語 B の用例作成を行う．また，言語 B 話者を複数人用いて，言語 B の用例が複数作成されるようにする．このとき，正しい用例の場合は複数人が作成すると考えられる．このため，本手法では複数人が全く同じ用例を作成した場合，その用例は他の用例より正確であると判断することとする．

[1]用例対訳とは，用例を多言語に正確に翻訳したコーパスのことを指す．

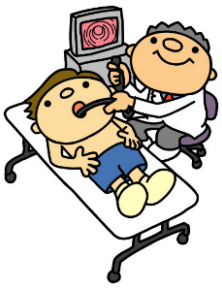

図 2: 実験で使用した画像の例
（参照：http://design.taiho.co.jp/）

Step 4 正確性評価
本手法では，単言語話者のみで用例対訳の作成を行っている．このため，実際に作成された用例対訳を使用するときは，図 1(4) のように，両言語を理解する多言語話者が確認する必要がある．

なお，本稿では元言語（言語 A）を日本語，翻訳先言語（言語 B）を英語として，手法の有用性確認を行う．また，英語話者はクラウドソーシング上で作業委託を行った．

4 多言語用例作成実験

4.1 用例の作成と翻訳

本節では，図 1(1) で示した用例の作成について述べる．用例の作成は下記の手順で行った．

1. 画像を用いた用例作成
   日本人用例作成者 2 名に医療に関係する画像（イラスト）を提示し，その画像をもとに日本語用例を作成するよう依頼した．図 2 に，実験で使用した画像の例を示す[2]．なお，日本人用例作成者は多言語用例対訳共有システム TackPad[9] で活発に用例登録を行っていた日本人用例作成者から選定した．その際，各作成者に 10 個の画像を提示し，それぞれの画像に対して 5 文ずつ用例を作成するよう依頼した．この結果，用例作成者 1 名につき 50 文の用例を収集し，計 100 文の日本語用例を得た．なお，平均文字長は 12.0 文字，標準偏差は 3.26 文字であった．
2. 用例の正確性評価
   日本人用例評価者 3 名に，(1) で用例作成者が作成した日本語用例 100 文の正確性評価を依頼した．また，不正確なものは正しい文を記入するように依頼した．その後，評価結果をもとに，著者の一人が用例の修正を行った．
3. 翻訳
   日本語用例 100 文を機械翻訳でそれぞれ英語に翻訳を行った（図 1(2)）．以降，機械翻訳で翻訳した文を「翻訳文」とする．本稿では，機械翻訳エンジンとして言語グリッド [4] の J-Server を利用した．

[2]http://design.taiho.co.jp/の画像を利用した．

Base Sentence: Please see an X-ray.

How do you judge the fluency of "Base Sentence"? (required)
- Incomprehensible
- Disfluent English
- Non-native English
- Good English
- Flawless English

Please input Flawless English sentence to be based on "Base Sentence". (required)

図 3: タスクの例

以降，機械翻訳による対訳作成を「既存手法 1」とし，日本語用例と翻訳した英文の 100 対を提案手法と比較を行う基準とする．

4.2 対訳の作成

本節では，図 1(3) で示した対訳の作成について述べる．本稿は，英語の対訳作成の場として，クラウドソーシングを用いた．なお，クラウドソーシングサービスとして CrowdFlower[3]を介して Amazon Mechanical Turk[4]を利用した．また，用例の作成（図 1(1)）で使用した画像を提示するグループと提示しないグループの 2 つに分けて作業依頼を行っている．以降，画像を提示したグループの結果を「提案手法」，画像を提示しなかったグループの結果を「既存手法 2」とする．また，作成された文を「英文」と表記する．

以下に，クラウドソーシングで行ったタスクを示す．また，タスクの画面例を図 3 に示す．

1. 前節で作成した機械翻訳文を提示し，5 段階で機械翻訳文の流暢性評価[5]を依頼した．
2. 正しい英文の記入を依頼した．ただし，提示した機械翻訳文が正しいと判断した場合は類似文の記入を依頼した．

なお，1 対の評価と対訳作成につき 5 セント[6]支払った[7]．また，タスクは両グループとも機械翻訳文 100 文の評価を 10 人ずつ行うようにした．

作業依頼の結果，提案手法は 29 名から，既存手法 2 は 33 名から，それぞれ 1000 文の評価を取得した．この中で，明らかに不適切な英文を記入しているデータ[8]を除き，提案手法（画像提示あり）は 885 文，既存手法 2（画像提示なし）は 973 文の結果を得た．なお，既存手法 2（画像提示なし）の 71 文は英文の取得

[3]http://crowdflower.com/
[4]https://www.mturk.com/
[5]文献 [6] の評価基準を用いた．評価段階は，1: Incomprehensible, 2: Disfluent English, 3: Non-native English, 4: Good English, 5: Flawless English，である．
[6]約 4.5 円（2013/1/13 現在．1 ドル= 89.16 円で計算）．
[7]クラウドソーシングを用いた研究では，コストについても議論を行う場合が多いため，従来研究にならって金額を記載している．
[8]“Base Sentence”, “This is OK”, “understandable”など．

表 1: 作成された翻訳文と英文の例

用例作成者	機械翻訳			提案手法 (画像あり)			
日本語用例	翻訳文	評価	正確性	英文	人数	評価	正確性
こちらの影が見えますか？	Is this shadow seen?	3.33	×	Do you see this shadow?	3 人	4.67	○
				Do you see a shadow in the x-ray?	1 人	3.67	×
				Can you see this shadow?	1 人	5.00	○

・日本語用例は，医療従事者が患者に対してレントゲン写真を示しているイラストをもとに作成された．
・表中の評価は，翻訳者 3 名が行った日本語用例と翻訳文もしくは英文の適合性評価の平均である．
・表中の正確性は，適合性評価をもとに判定し，○が正確，×が不正確を示す．
・この翻訳文に対しては 8 名の労働者が評価した．そのうち，3 名が流暢性評価で 4 と評価した．表中の提案手法には流暢性評価で 3 以下をつけた労働者の結果のみを表示している．

に失敗したため，次章の分析で英文を用いるものに関しては，902 文のみを用いる．

4.3 用例対訳候補の正確性判定

本節では，図 1(4) で示した用例対訳候補の正確性判定について述べる．本稿では，本節の評価結果を各手法の評価に用いる．

本稿では，翻訳者 3 名に 5 段階で日英対の適合性評価[9]を依頼した．評価依頼を行った日英対は下記のものである．

- 既存手法 1（機械翻訳）で作成された機械翻訳文（英文）と，もとの日本語文の対（100 対）．
- 流暢性評価で 3 以下と判断された機械翻訳文に対して，提案手法（画像提示あり）で作成された英文と，もとの日本語文の対（重複を除いた 444 対）．
- 流暢性評価で 3 以下と判断された機械翻訳文に対して，既存手法 2（画像提示なし）で作成された英文と，もとの日本語文の対（重複を除いた 352 対）．

なお，適合性評価の平均が 4 より大きいものを正確，4 以下のものを不正確と判定し，次章の分析で用いた．

5 考察

5.1 各手法の正確性

本節では，各手法の正確性について考察する．

まず，表 1 に実験で作成された翻訳文と英文の例を示す．なお，この文は，医療従事者が患者に対してレントゲン写真を示しているイラストをもとに作成された（図 3）．表 1 は機械翻訳で不正確な日英対となっていたものが，提案手法によって正確な日英対となった例を示している．

表 2 に既存手法 1（機械翻訳）の翻訳精度を示す．表 2 より，既存手法で作成された英文は，8 割以上がそのままでは使用できないことが分かる．このため，用例対訳作成においては，機械翻訳をそのまま使用することは適切ではないことが分かる．

次に，表 3 に提案手法（画像提示あり）と既存手法 2（画像提示なし）の判定精度を示す．なお，労働者の判定は，労働者の流暢性評価の平均が 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ている．

[9]文献 [6] の評価基準を用いた．評価段階は，「1：全く違う意味」「2：雰囲気は残っているが元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．

表 2: 既存手法 1（機械翻訳）の翻訳精度

	不正確	正確	合計
評価数	86	14	100

・翻訳者 3 名の 5 段階評価の平均が，4 より大きいものを正確，4 以下を不正確と判定している．

表 3: 提案手法と既存手法 2 の判定精度

労働者の判定	正確と判定		不正確と判定		正解率
翻訳者の判定	正確	不正確	正確	不正確	
提案手法 (画像あり)	10	7	4	79	89%
既存手法 2(画像なし)	12	14	2	72	84%

・労働者の正解率は，労働者の判定と翻訳者の判定が同一となった割合を示す．
・合計は 100 文である．

表 3 より，提案手法，既存手法 2 ともに，比較的正確に判定できていることが分かる．しかし，翻訳者が不正確と判定したものを，労働者が正確と判定した例が，提案手法では 7 文，既存手法 2 では 14 文存在していた．これは，労働者と翻訳者はそれぞれ流暢性評価と適合性評価を行っており，評価軸が異なっていたことが原因であると考えられる．流暢性評価（労働者）で正確，適合性評価（翻訳者）で不正確と判定された文として，「It's a cold.」がある．この文は，原文の「風邪です」とはあまり適合していないが，英語の流暢性は高い．このため，労働者は提示された画像に適した正しい対訳を作成できなかったと考えられる．このため，画像の提示を行っている提案手法では，流暢性評価の他に，英文が画像に適しているかどうかを評価する必要があると考えられる．

5.2 画像の提示有無による正確性

本節では，画像の提示の有無による用例対訳の正確性について考察する．なお，本稿では，労働者の流暢性評価で不正確であると判定された機械翻訳文（提案手法：83 文，既存手法 2：74 文）から作成された英文を，提案手法と既存手法 2 の比較で用いる．その際，流暢性評価で 3 以下と評価された，提案手法の 524 文と，既存手法 2 の 408 文をそれぞれ用いた．なお，これらの文数は重複を含んだ数である．

5.2.1 提案手法の正確性

表 4 に各手法の正確割合を示す．表 4 は，作成されたすべての英文（提案手法：885 文，既存手法 2：902 文）をもとに調査を行った．なお，正確，不正確の判定は，翻訳者による適合性評価を用いている．また，

表 4: 各手法の正確割合

	評価利用基準		
	1 名以上	2 名以上	3 名以上
提案手法 (画像あり)	36.9% (83)	50.0% (66)	69.7% (33)
既存手法 2(画像なし)	27.6% (74)	37.3% (51)	52.2% (23)

・評価利用基準の人数は，同じ文を記入した労働者の閾値である．
・表中の割合は，各手法で作成された英文 の正確割合である．分析対象の英文数は，提案手法が 885 文，既存手法 2 が 902 文である．
・表中の括弧内は，使用している翻訳文 の数である．分析対象の翻訳文数は，提案手法が 83 文，既存手法 2 が 74 文である．

表 5: 翻訳文の適合性評価と日英対の正確性

翻訳文の	提案手法の英文	
適合性評価平均	正確率	対象文数
1.00	0.0%	8
1.33	6.1%	33
1.67	8.7%	23
2.00	5.7%	53
2.33	16.3%	43
2.67	30.0%	90
3.00	39.1%	87
3.33	42.7%	110
3.67	78.8%	52
4.00	56.0%	25

表 4 中の評価利用基準の人数は，同じ文を記入した労働者の閾値である．

表 4 より，画像の提示を行っている提案手法は，画像の提示を行っていない既存手法 2 より 9.3 ポイント (表 4 の評価利用基準が 1 名以上との差)，元の日本語用例と正しい対となる英語用例の作成ができていることが分かる．また，評価利用基準を 2 名以上や 3 名以上と制限することで，正確な日英対が作られる割合が大きくなっていることが分かる．複数人が作成した英文は，正しい多言語対の作成に貢献していると考えられる．

5.3 機械翻訳の精度と日英対

本節では，機械翻訳の精度と，それをもとに作成される英文の関係について考察する．表 5 に機械翻訳で翻訳した文の適合性評価と，提案手法で作成された日英対の正確性の関係を示す．表 5 より，機械翻訳の適合性評価の平均が高い翻訳文の方が，適切な英文作成の手がかりとなっていることが分かる．

なお，適合性評価平均が 4.00 のときに正解率が下がっているのは，翻訳文が「Medicine for 3 days.」(原文は「3 日分のお薬です」) をもとに作成された英文が，適合性評価ですべて不正確であったことが要因となっていた．

5.4 クラウドソーシングのコスト

本節では，クラウドソーシングを用いることによるコストについて議論する．

翻訳者による翻訳業務を行っているエキサイト翻訳依頼プロ[10]の場合，医療用例の翻訳には 1 文字あたり 25 円が必要である[11]．今回利用した用例は平均 12.0 文字であった．このため，1 文の翻訳に約 300 円かかることが分かる．

提案手法では，1 文あたり 50 セント（=5 セント× 10 人）で対訳の作成を行う．また，1 文あたり 50 セント（=5 セント× 10 人）で多言語対の評価を行っている [7]．これらから，約 90 円[12]で翻訳が可能となる．

さらに，翻訳者による翻訳は対訳が 1 文のみ生成されるが，本手法では複数の対訳が生成される．多言語用例対訳は多様性を持つことで自由度を高めた多言語変換を行うことが可能である [8]．これらのことから，安価で多様性を持つ多言語対の作成が可能な本手法は有用であると考えられる．

6 おわりに

本稿では，単言語話者のみで多言語用例対訳候補の作成を行う手法の提案とその実験を行った．本手法では，画像を単言語話者に提示することで，正確な多言語対の作成を目指した．

本稿の貢献は，画像を用いることで，単言語話者のみで機械翻訳よりも高精度の多言語用例対訳候補の作成が可能であることを示した点である．

今後は，他のデータセットを用いて同様の結果が得られるか追加実験を行う．その際，画像と英文の適切性について確認を行う．

参考文献

[1] Chris Callison-Burch. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing(EMNLP2009)*, pp. 286–295, 2009.

[2] Anhai Doan, Raghu Ramakrishnan, and Alon Y. Halevy. Crowdsourcing systems on the world-wide web. Vol. 54, No. 4, pp. 86–96, 2011.

[3] Jeff Howe. Crowdsourcing: Why the power of the crowd is driving the future of business. Crown Business, 2008.

[4] Toru Ishida. Language grid: An infrastructure for intercultural collaboration. In *IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06)*, pp. 96–100, 2006.

[5] Matteo Negri and Yashar Mehdad. Creating a bi-lingual entailment corpus through translations with mechanical turk: $100 for a 10-day rush. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 212–216, 2010.

[6] Kevin Walker, Moussa Bamba, David Miller, Xiaoyi Ma, Chris Cieri, and George Doddington. Multiple-translation arabic (mta) part 1. In *Linguistic Data Consortium, Philadelphia*, 2003.

[7] 福島拓, 吉野孝. 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. Web とデータベースに関するフォーラム (WebDB Forum 2012), No. B4-3, 2012.

[8] 福島拓, 吉野孝. 正確な多言語間対話支援を目的とした応答用例対構築モデルの検討. 情報処理学会，マルチメディア，分散，協調とモバイル（DICOMO2012）シンポジウム, pp. 551–559, 2012.

[9] 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.

[10] https://orderpro.excite.co.jp/

[11] 2012/12/12 現在.

[12] 2013/1/13 現在．1 ドル= 89.16 円で計算.