E-08

後編集文と機械翻訳文との意味の近さを用いた単言語話者による用例対訳作成手法の提案

Proposal of Parallel-text Creation Method by Monolingual People Using Closeness of Meaning Between Post-editing Text and Machine Translation Text

山本　里美 †　　福島　拓 ‡　　吉野　孝 †
Satomi Yamamoto　Taku Fukushima　Takashi Yoshino

1. はじめに

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は増加している．多言語間コミュニケーションの支援として，機械翻訳や用例対訳などが用いられている．正確な情報の共有が重要となる医療分野などでは，十分に正確性の確保された用例対訳が使用されている．しかし，正確な情報共有のために必要とされる用例対訳の数は多く，十分な数の用例対訳の収集は困難である [1]．我々は，現在，機械翻訳文をクラウドソーシングにおいて評価・訂正を依頼することで，単言語話者であっても用例対訳作成を行うことができる手法の研究を行っている．これまで，クラウドソーシング作業者に会話文の形式で機械翻訳文を提示することで，用例対訳や応答用例対 [*1] を作成できることが分かった [2]．また，文献 [2] では，翻訳元の言語への機械翻訳を用いて，クラウドソーシングによって日本人作業者に正確性評価を依頼する手法の提案と評価実験を行っている．評価実験の結果，専門家による正確性評価の結果の一部と一致することが分かった．

しかし，文献 [2] の評価実験では，作成された用例対訳候補文のうち，専門家が正確な対訳と評価した文は 110 文だったが [*2]，クラウドソーシングを用いた正確性評価で，正確だと判断された文は 14 文だった．なお，用例対訳候補文とは，クラウドソーシングの作業者によって高い評価がなされた機械翻訳文や，機械翻訳文の訂正文（後編集文）のことである．この 14 文は，作業者が正確だと評価した文と一致した 110 文に含まれる文だったが，クラウドソーシングでは残りの 96 文を取得することはできなかった．このことから，クラウドソーシングを用いた日本人作業者による正確性評価では，正確な対訳を取得することはできるが，その抽出率が非常に低いことが分かった．

本稿では，日本人作業者による正確性評価ではなく，翻訳対象言語を母語とする作業者に，機械翻訳文と後編集文を提示して行う正確性評価手法を提案し，評価実験を行った．機械翻訳文と後編集文を提示することで，作業者は機械翻訳文から元の用例（機械翻訳前の用例）の意図を推測し，それにあった後編集文を選択する可能性があると考えたためである．

2. 関連研究

現在，クラウドソーシングを用いて多言語データを収集する研究が多く行われている．多言語テキストの正確性評価をクラウドソーシングの作業者に依頼する研究や [3]，翻訳対象の文やクラウドソーシング作業者の特徴をもとに，翻訳文を分析し，品質の良い翻訳文を作成する研究 [4]，クラウドソーシングと機械翻訳を併用してより品質の高い対訳コーパスや翻訳結果を取得する研究 [5, 6] などがある．多くの研究において，クラウドソーシングの作業者は非専門家を対象としているが，多言語を理解することができることが前提である．そのため，英語や中国語など，話者の多い言語についての対訳コーパス作成においては，多くのデータを収集可能だが，話者の少ない言語では，その言語を理解できる作業者が少ないため，十分な数の対訳コーパスの作成は困難であると考えられる．

そこで我々は，クラウドソーシング上に多くいる単言語話者による用例対訳作成手法について研究を行っている．機械翻訳後の言語を母語とする作業者に機械翻訳文を提示した場合，それが正確性の低い機械翻訳文であっても，翻訳前の文の意図を推測し，元々の意図にあった訂正を行うことが可能なのではないかと考えたためである．なお，実験の結果，一部の用例においては単言語話者による用例対訳作成が可能であることがわかった [7]．用例対訳として使用するためには正確性確保のための評価が必要である．しかし，大量の用例対訳の正確性評価を専門家に依頼した場合，人的リソースやコストの面で問題がある．本稿では，クラウドソーシングを用い，翻訳後の言語を母語とする作業者によって正確性評価が行えるかどうかの検証を行う．

3. クラウドソーシングを用いた用例対訳作成手法と従来の正確性評価手法

本章では，文献 [2] における，クラウドソーシングを用いた，単言語話者による用例対訳作成手法と正確性評価手法について述べる．図 1 に用例対訳作成と従来の正確性評価の流れを示す．

3.1 用例対訳作成手法

文献 [2] の用例対訳作成手法では，疑問文とその回答，それに続く 1 文の，3 文からなる会話文を使用する．会話文を提示することで，図 1(1) の機械翻訳文の訂正を行う作業者は，文脈から翻訳前の用例の意図を推測しやすくなり，訂正精度が向上すると考えたためである．ただし，会話文を提示する場合，文脈に合わせた訂正が行われるため文の流れは変更しないが，翻訳前の用例とは意味の異なる後編集文が作成される可能性がある．なお，このような後編集文は，応答用例対として使用可能である．文献 [2] の手法では，図 1(2) のように，機械翻訳文の流暢性評価の評価値を用いて分類することで，用例対訳と応答用例対の分類を機械的に行う．流暢性評価の基準は，文献 [8] の評価基準 [*3] を参考にしている．従来手法では，文法などの他に，会話の流れとして適切かどうかも加味して評価を行う．

† 和歌山大学，Wakayama University
‡ 静岡大学，Shizuoka University

[*1] 質問とその回答の対，またその類似文からなる用例対訳のことであり，会話の支援に用いられる．

[*2] 3 名の作業者のうち 1 人でも正確と判断した場合の数．

[*3] 評価段階は，1 : Incomprehensible(理解できない)，2 : Disfluent English(流暢でない)，3 : Non-native English(非母語言語)，4 : Good English(良い英語)，5 : Flawless English(完璧な英語)

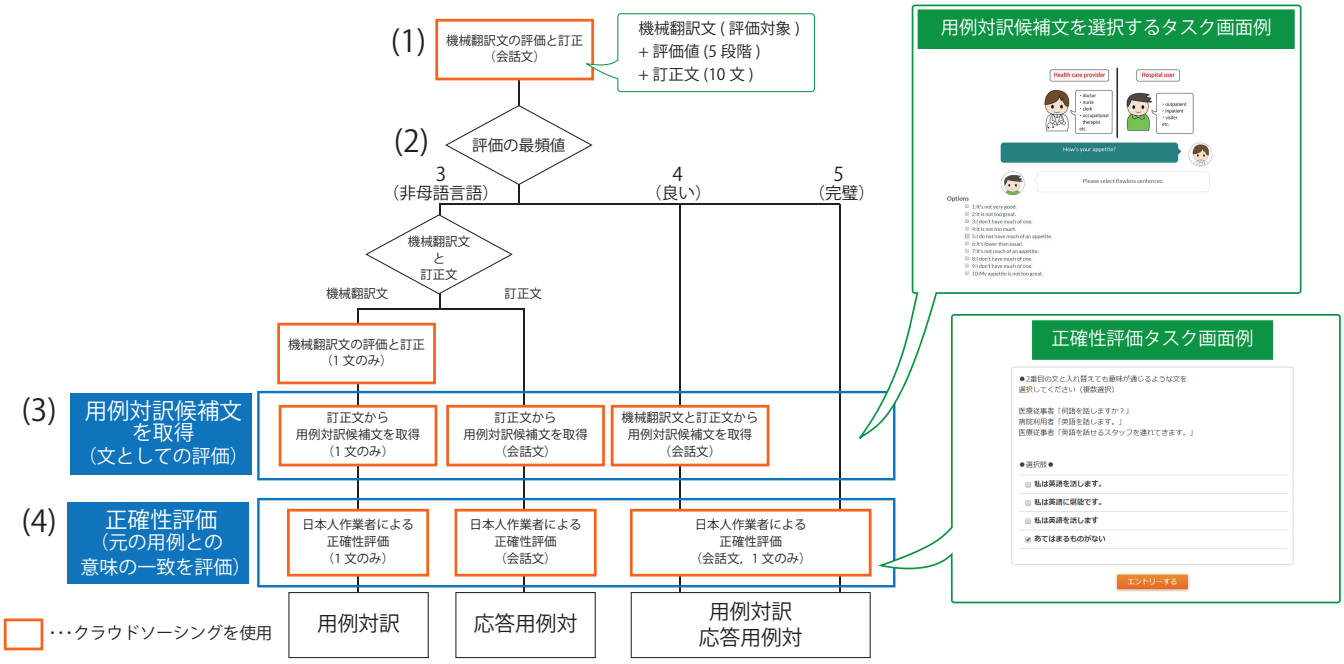

図 1: 用例対訳作成と従来の正確性評価の流れ

分類後，図 1(3) で，翻訳後の言語を母語とする作業者に用例対訳候補文を選択するタスクを依頼し，用例対訳候補文を取得する．用例対訳候補文を選択するタスクでは，作業者には，機械翻訳文と訂正文を選択肢として提示し，医療従事者または病院利用者が使用する上で流暢 (Flawless) な文を選択してもらう．なお，図 1 で（会話文）とある場合のタスクでは，作業者に選択肢にある文は会話文中のものであることを示す．

3.2 従来の正確性評価手法

図 1(4) で，取得した用例対訳候補文の正確性評価は，クラウドソーシングを用いて，翻訳元の言語を母語とする作業者に依頼する．用例対訳候補文を元の用例の言語に機械翻訳し，作業者には元の用例と似た意味を示す機械翻訳文を選択するタスクを依頼する．

3.3 従来の正確性評価手法を用いた用例対訳作成実験

本節では，図 1 で示す手法を用いた用例対訳作成実験の結果について述べる．使用した会話文は 38 組である．機械翻訳文の評価と訂正タスク，用例対訳候補文選択タスク，正確性評価タスクの 3 種類のタスクがあるが，すべてのタスクにおいて，1 文あたり 10 名の作業者に評価作業や選択作業を依頼している．また，日本語の用例から英語の用例対訳作成を行う．実験の結果，用例対訳と使用可能と考えられる用例対訳候補文は 14 文，応答用例対として使用可能と考えられる用例対訳候補文は 26 文であった．用例対訳として使用可能と考えられる用例対訳候補文 14 文は，専門家に依頼した正確性評価によって正確性が高いと判断された用例対訳候補文と一致していた．そのため，クラウドソーシングによる正確性評価手法は有用であると考えられる．しかし，専門家によって有用と判断された用例対訳候補文の総数は 110 文であり，クラウドソーシングによる正確性評価手法で取得できた用例対訳は，全体の 12%であることがわかった．よって，用例対訳候補文から正確性の高い用例対訳の抽出率を向上させる必要がある．

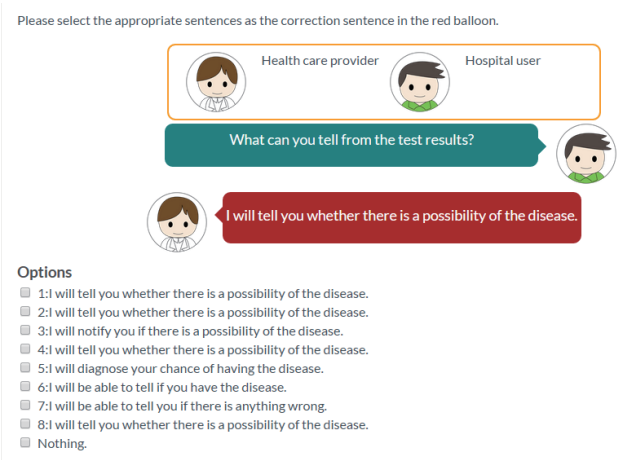

図 2: 提案手法における用例対訳候補文選択タスク画面例

4. 提案手法

従来の正確性評価手法では，正確性の高い用例対訳を取得することはできるが，抽出率が低いことがわかった．これは，翻訳元の言語への機械翻訳における翻訳精度が影響している可能性がある．そこで，本稿では，翻訳元の言語への機械翻訳を用いない正確性評価手法を提案する．本手法では，図 1(1)，(2) は従来手法と同様に行うが，図 1(3) の用例対訳候補文取得の段階で，図 1(4) で行っていた正確性評価（元の用例と意味が一致しているかどうかの評価）を同時に行う．従来手法では，用例対訳候補文選択の際に，作業者に訂正文の元となる機械翻訳文は提示していなかった．提案手法における用例対訳候補文選択タスク画面の例を図 2 に示す．なお，評価値 3 に分類された機械翻訳文は 1 文のみ提示での評価と訂正を行う．図 2 は会話文を提示する場合のタスク画面の例である．従来の正確性評価手法を用いる場合の用例対訳候補文選択タスクでは，赤い吹き出しには “Please select flawless sentences.”と書いていた．提案手法では，訂正文の元となった機械翻訳文を提示する．用例対

表 1: 評価実験で用いた作業者への指示

指示	タスク全体の説明文	各設問における指示文
1	The sentence in the red balloon is the sentence made by machine translation. Please select the same meaning sentences at the list below.	Please select the same meaning sentences as the sentence in the red balloon.
2	The sentence in the red balloon is the sentence made by machine translation. The sentences at the list below are the candidates of the correction sentence. Please select the appropriate sentences as a correction sentence.	Please select the appropriate sentences as the correction sentence in the red balloon.

・指示 1 は，機械翻訳文と同じ意味を示す文を選択するように指示している．
・指示 2 は，機械翻訳文の訂正文として適切な文を選択するように指示している．

表 2: 各タスクにおいて取得した文と専門家による評価と一致した文の数

手法		用例対訳候補文選択	正確性評価	専門家による評価と一致
従来手法		278	14	14
提案手法	指示 1	89		19
提案手法	指示 2	17		2

・提案手法では用例対訳候補文選択と正確性評価を同時に行っている．

訳候補文選択の際に，作業者に機械翻訳文を提示することで，機械翻訳文と訂正文を比較し，文脈を推測する可能性があり，元の用例との意味の一致まで行えるのではないかと考えた．

5. 提案手法を用いた正確性評価実験

文献 [2] で行った用例対訳作成実験と同様のデータを用い，提案手法の評価実験を行った．本実験では，用例対訳候補文選択タスクにおいて，作業者に 2 通りの指示を出した．それぞれの指示を表 1 に示す．タスク全体の説明文は，タスクページの上部に 1 度だけ表示される文章であり，各設問における指示文は，設問 1 つにつき 1 度表示される文章である．なお，本実験では，6 つの設問への回答を 1 タスクとしている．表 1 の指示 1 は,「同じ意味」かどうかについての評価，指示 2 は,「訂正文として適切」かどうかについての評価を依頼している．また，どちらの場合も，タスク画面における赤い吹き出し部分にある文が機械翻訳文であることを提示している．

従来手法と，提案手法における各タスクで取得した文の数と，専門家による評価と一致した文の数を表 2 に示す．従来手法では，用例対訳候補文選択と正確性評価は別のタスクとしていたが，提案手法では，用例対訳候補文選択を行う際に，意味の一致が考慮される可能性があることから，用例対訳候補文選択と正確性評価を 1 つのタスクで行っている．表 2 より，従来手法で正確と判断された用例対訳候補文の数は専門家による評価とすべて一致している．しかし，提案手法では，どちらの指示の場合も，正確性評価で選択された文の数のうち，専門家による評価と一致した文の数は，指示 1 の場合は 89 文中 19 文，指示 2 の場合は 17 文中 2 文と，従来手法と比べて少ないことが分かる．用例対訳として使用する際に必須とされる正確性の確保は，専門家の評価と一致している必要があるため，取得した用例対訳候補文の数と，専門家による評価と一致した文の数は同じである必要がある．

本実験において，提案手法で専門家による評価との一致率が低かった理由として，誤った機械翻訳文が提示された場合に，作業者が同じ意味の文を選択しようとしたことや，逆に，機械翻訳文の意味を考慮せず，会話文として成立する文を選択しようとしたことが考えられる．例えば，元の機械翻訳文 “Let’s watch over the status quo.”であるのに対し，“Huh?”という訂正文が選択される場合があった．これは，会話文に登場する文としては適切であると作業者が判断したためだと考えられる．日本人作業者による正確性評価では，機械翻訳前の正しい日本語と，訂正文の日本語への機械翻訳文を比較することで正確性評価を行っているが，提案手法では，精度の低い機械翻訳文と，その訂正文が提示されているため，作業者の推測が大きく影響する可能性がある．これによって，本実験のような一致率の低い正確性評価の結果になったと考えられる．

6. おわりに

本稿では，クラウドソーシングを用いた単言語話者による用例対訳作成手法について説明し，従来手法において行っていた，翻訳元の言語への機械翻訳を用いない正確性評価手法を提案し，評価実験を行った．評価実験より，翻訳後の言語を母語とする作業者のみでの用例対訳の正確性評価は，従来手法である元言語への機械翻訳を行う正確性評価よりも，精度が低いことがわかった．今後は，元言語への機械翻訳を用いた用例対訳の正確性評価手法において，抽出される用例対訳候補文の数を増やすとともに，専門家による評価との一致率の向上を目指す．

参考文献

[1] 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).

[2] 山本 里美, 福島 拓, 吉野 孝: クラウドソーシングにおける機械翻訳文の評価結果を活用した用例対訳作成手法の提案, 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.2015-GN-93, No.38, pp.1-8 (2015).

[3] Callison-Burch, C.: Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon’s Mechanical Turk, *Proceedings of EMNLP 2009*, pp. 286–295 (2009).

[4] F.Zaidan, O. and Callison-Burch, C.: Crowdsourcing Translation:Professional Quality from Non-Professionals, ‘11 Proceedings of the 49th Annual Meeting of the Association for

Computational Linguistics: Human Language Technologies, Vol.1, pp. 1220–1229 (2011).

[5] Vamshi Ambati，Stephan Vogel : Can Crowds Build Parallel Corpora for Machine Translation Systems?，Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk，pp. 62–65 (2010).

[6] Philip Resnik，Olivia Buzec，Yakov Kronrod，Chang Hu，Alexander J. Quinn，Benjamin B. Bederson: Using Targeted Paraphrasing and Monolingual Crowdsourcing to Improve Translation, Transactions on Intelligent Systems and Technology (TIST)，Vol.4，No.3，Article No.38 (2013).

[7] 福島 拓, 吉野 孝: 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, 言語処理学会第 19 回年次大会 (NLP2013), pp. 302–305 (2013).

[8] Kevin Walker，Moussa Bamba，David Miller，Xisoyi Ma，Chris Cieri，and George Doddington: Multiple-Translation Arabic (MTA) Part 1，In Linguistic Data Consortium，Philadelphia (2003).