情報処理学会論文誌 Vol.56 No.3 1080–1089 (Mar. 2015)

クラウドソーシングにおける会話文を用いた応答用例対作成手法の提案

山本 里美[1,a)] 福島 拓[2,b)] 吉野 孝[1,c)]

受付日 2014年6月30日, 採録日 2014年12月3日

概要：現在，グローバル化によって多言語間コミュニケーションの機会が増加している．正確な情報の共有が重要となる医療分野などでは，コミュニケーション支援の方法として，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，必要な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．本論文では，正確性の低い機械翻訳文を，クラウドソーシングにおいて訂正を依頼することで，単言語話者でも応答用例対の作成が行える手法を提案する．提案手法は，作業者が用例の意図を推測しやすくするために，評価対象の機械翻訳文を含んだ「会話文」を提示する．本論文の貢献は，応答用例対作成における会話文の提示の効果を明らかにした点である．

キーワード：クラウドソーシング，多言語間コミュニケーション，用例対訳，機械翻訳

Proposal of a Pair of Dialogic Parallel Texts Creation Method Using Conversation Sentences on Crowdsourcing

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

Received: June 30, 2014, Accepted: December 3, 2014

Abstract: The opportunities for communication amongst people whose native languages differ are increasing by globalization. However, it is difficult to share exact information among many languages. A parallel text that combines example sentences and their accurate translation are used in a multilingual communication support. There are many required parallel texts. Moreover, it is not easy to collect a sufficient number of parallel texts. In this paper, we propose a pair of dialogic parallel text creation method of requesting correction for a machine translation sentence with low accuracy on crowdsourcing of a monolingual. The proposed method presents the conversation sentences with a machine translation sentence, in order that a worker may make the intention of example sentences easy to guess. The contribution of this paper is to clarify the effect of presentation of the conversation sentence in a pair of dialogic parallel text creation.

Keywords: crowdsourcing, multilingual communication, parallel text, machine translation

1. はじめに

現在，世界的なグローバル化によって，多言語間コミュニケーションの機会が増加している．日本でも，外国人留学生数の増加 [1] や，2013 年の訪日外国人数が 1,000 万人を突破したこと [2]，日本政府が 2020 年に訪日外国人数 2,500 万人という目標を掲げていること [3] から，今後も訪日外国人数は増加すると考えられる．そのため，今後の多言語間コミュニケーションの機会も増加していくと考えられる．しかし，一般的に多言語を十分に習得することは難しく，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を共有することは非常に困難である．

その影響が顕著に現れる分野の 1 つとして医療分野がある．医療分野では，わずかなコミュニケーション不足により医療ミスが生じる恐れがあり，正確な情報の共有が非常

[1] 和歌山大学システム工学部
Faculty of Systems Engineering, Wakayama University, Wakayama 640–8510, Japan
[2] 静岡大学大学院工学研究科
Graduate School of Engineering, Shizuoka University, Hamamatsu, Shizuoka 432–8561, Japan
a) s165058@center.wakayama-u.ac.jp
b) fukushima@sys.eng.shizuoka.ac.jp
c) yoshino@sys.wakayama-u.ac.jp

に重要となっている．現在，医療現場において日本語の理解できない外国人に対する支援は主に医療通訳者が行っている．しかし，慢性的な人員不足や，通訳者の身分保障，通訳者自身のメンタルケアなどの問題が存在している．また，日本語の理解できない外国人が入院する場合，24 時間の対応が必要となり，医療通訳者への負担が大きくなると考えられる．このような問題は，外国人が多くない地域でも存在しているため，インターネットを用いた多言語間コミュニケーション支援として，アイコンを用いたコミュニケーション支援システムの開発 [4] や，用例対訳や機械翻訳などの言語資源を組み合わせて利用することのできる言語グリッド[*1]の活動などが行われている [5]．

現在，我々は，多言語用例対訳共有システム TackPad（タックパッド）[*2]による用例対訳の収集を行っている [6]．TackPad には多くの用例や用例対訳[*3]が収集されているが，医療現場で用いるために必要とされる数[*4]には足りておらず，また，正確性評価が十分にされていないものも多い．専門家の代わりに，クラウドソーシングを用いて不特定多数の人に業務を委託する方法がある．しかし，機械翻訳文の評価や訂正は文脈に依存するため，正確性の高い評価や文脈に沿った訂正文を得られないという問題がある．

そこで本論文では，クラウドソーシングを用いて会話文を作成し，機械翻訳によって別の言語に翻訳した後に，クラウドソーシングによって評価と修正を依頼することで，専門家に依頼することなく応答用例対の作成を行う手法を提案する．本手法で作成する会話文は，質問を会話の始点とし，それに対する回答，その回答に続く返答文の 3 文で構成されるものとする．また，会話文の始点となる質問は，すでに TackPad に用例対訳が存在する用例とする．なお，応答用例対は文献 [7] での定義に従い，1 個の質問と 1 個の回答，0 個以上の類似文から構成されるものとする．

2. 関連研究

現在，多言語間コミュニケーション支援を目的とした，機械翻訳や用例対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章を多言語に翻訳することが可能なため，単言語対応のチャットシステム [8] や，使用言語の異なる複数人グループでのコミュニケーション支援 [9] などで利用されている．また，機械翻訳の翻訳精度を向上させる研究は多く行われているが，人手による操作や編集を用いることで翻訳精度の向上を目指す手法として，前処理や後編集，翻訳リペアなどの研究 [10], [11], [12] があり，誤りがある機械翻訳文であってもコミュニケーションを円滑に行うことを可能とするための手法の研究 [13] など，機械翻訳に関する研究が多く行われている．機械翻訳の精度は年々向上しているが，現在の翻訳精度では，正確な情報の共有が重要となる医療分野で用いることができるだけの正確性を確保することは困難である．そのため，医療分野などの正確性が重要となる分野でのコミュニケーション支援では，用例対訳を用いた方法が多く行われている．用例対訳を用いたシステムには，多言語医療受付支援システム M^3（エムキューブ）[14] や，入院生活での多言語対話の支援を行うスマートフォン対応の多言語医療対話支援システムぷち通 [15] などがある．

また，計算機では判断や作成が困難なデータに対して，クラウドソーシングを用いることで，正確なデータを取得する研究が多く行われている．計算機で困難とされる，画像に含まれる情報の読み取りや，情報の収集などを行うことができるため，計算機を用いて情報の収集や判断を行うよりも正確な情報が得られる場合がある [16]．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おり，多言語話者による用例対訳作成 [17] や多言語テキストの正確性評価 [18] を行う研究などがある．また，クラウドソーシングと機械翻訳を併用することによって，専門家に依頼することなく，より品質の良い対訳コーパスや翻訳結果を取得しようとする研究が多く行われている [19], [20]．

我々は，正確性の低い機械翻訳文であっても，翻訳後の言語を母語とする人であれば，翻訳前の用例の意図を推測し，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．そこで，機械翻訳の翻訳後の言語を母語とするクラウドソーシング上の作業者に修正の依頼を行い，機械翻訳文の評価と訂正文を取得することによって，翻訳前の文と取得した訂正文から用例対訳を作成する手法を提案した [21]．単言語話者によって作成された訂正文と，機械翻訳前の用例を組み合わせることによって，用例対訳を作成することができれば，多くの用例が必要となる用例対訳の課題の 1 つである人的リソースの不足に対する解決方法の 1 つとなる可能性がある．しかし，すべての正確性の低い機械翻訳文を訂正することは困難である．そのため，評価と訂正を行うクラウドソーシングの作業者に提示する情報によって，訂正精度の向上をはかる研究を行っている [22]．本研究では，正しい用例対訳と，評価対象の機械翻訳文を含んだ会話文を用い，評価対象文が使用される状況を限定して作業者に提示することで，正確性の低い機械翻訳文を，正確性の高い会話文へ訂正することを目指す．

3. 提案手法

本章では，会話文を機械翻訳を用いて別の言語に翻訳し，翻訳文に対してクラウドソーシングを用いて評価・訂正を行うことで応答用例対を作成する手法について述べる．

[*1] http://langrid.org/jp/
[*2] http://med.tackpad.net/
[*3] 本論文では，正確性の確保が行われた多言語の用例の対を「用例対訳」，正確性の確保が行われていない多言語の対を「多言語テキストペア」とする．
[*4] 1 言語あたり 3 万〜5 万文 [6]

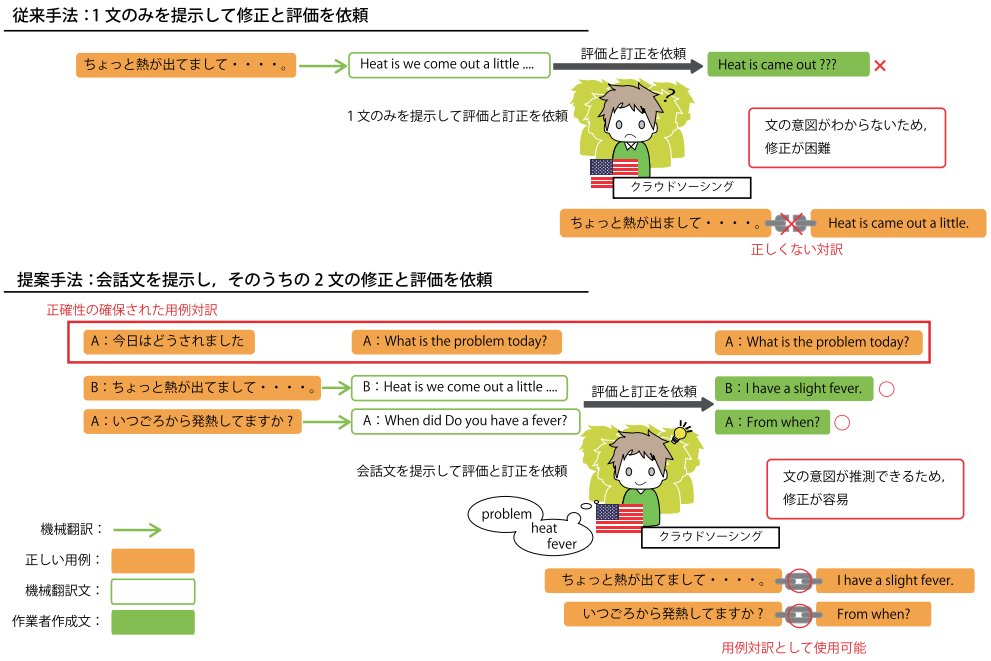

図 1　従来手法と提案手法の比較

Fig. 1　Comparison between a conventional method and a proposed method.

3.1　提案手法と従来手法との比較

本手法では，作業者に正しい用例対訳を含んだ会話文を提示することで，正確性の低い機械翻訳文を，翻訳前の用例の意図に沿った文へ訂正することを目指す．我々のこれまでの実験において，評価対象の機械翻訳文の訂正を行う際，“Pain was Zukinzukin will.” という機械翻訳文を “Pain was terrible.” のように，“Zukinzukin” を使用せず，代わりに “terrible” を用いて痛みについて説明をしたり，“Nail was peeling” という機械翻訳文に対して，“Her nail was peeling” のように “Her” を追加し，話しているのが誰の爪についてなのかを明確にしたりするなど，機械翻訳文に含まれていない単語を用い，機械翻訳文が使用される状況をより詳細に説明するような訂正を行う作業者がいた [22]．このような作業者が行った訂正は，翻訳前の文の意図としては不適切な場合もあったが，医療現場で使用する用例として利用可能であった．このため，不十分な翻訳がなされた機械翻訳文の訂正を行う場合，作業者は文の使用場面などを想像し，その想像した状況にあった訂正を行うと考えられる．そのため，我々は，作業者に評価対象の機械翻訳文を提示する際，その機械翻訳文を含む会話文を提示することで，従来手法でみられたように，それぞれの作業者が評価対象の機械翻訳文が使用される状況を想像するのではなく，その文の使用される状況をあらかじめ限定しておけば，用例の意図にあった訂正文の作成が行えるのではないかと考えた．

図 1 に従来手法（文献 [22]）と提案手法について示す．従来手法では，機械翻訳文の訂正を依頼する際，作業者には評価対象の翻訳文 1 文のみを提示する．そのため，正確性が低い機械翻訳文では元の文の意図を推測でない場合，適切な修正が困難となる可能性があった．提案手法では，作業者に評価対象の機械翻訳文を含む会話文を提示することで，文が使用される状況を従来手法より明確に示している．そのため，作業者は元の文の意図を推測しやすくなり，正確性が低い機械翻訳文でも適切な訂正が行える可能性があると考えた．

なお，本論文では，本手法の比較対象として，機械翻訳文の評価・訂正を行う際に，作業者に評価対象の機械翻訳文のみを提示する手法を従来手法とする．

3.2　提案手法の概要

本節では，提案手法の具体的な内容について説明する．**図 2** に提案手法の流れを示す．本手法は以下の 4 ステップで構成されている．

Step 1　会話文の作成

図 2 の (1) で，すでに正確性の確保された用例対訳の用例を会話の始点とし，会話文を作成する．

Step 2　機械翻訳文の取得

図 2 の (2) で，Step 1 で取得した会話文の質問に対する回答と，その回答に対する返答に対し，機械翻訳を用いて別の言語に翻訳する．ただし，TackPad にすでに用例対訳が存在しているときは，機械翻訳を行わず TackPad の用例対訳を用いる．これは，提示される会話文中に含まれる正確性の確保された文が増えることで，Step 3 の作業者が会話文の使用される状況をより想像しやすくなると考えたためである．

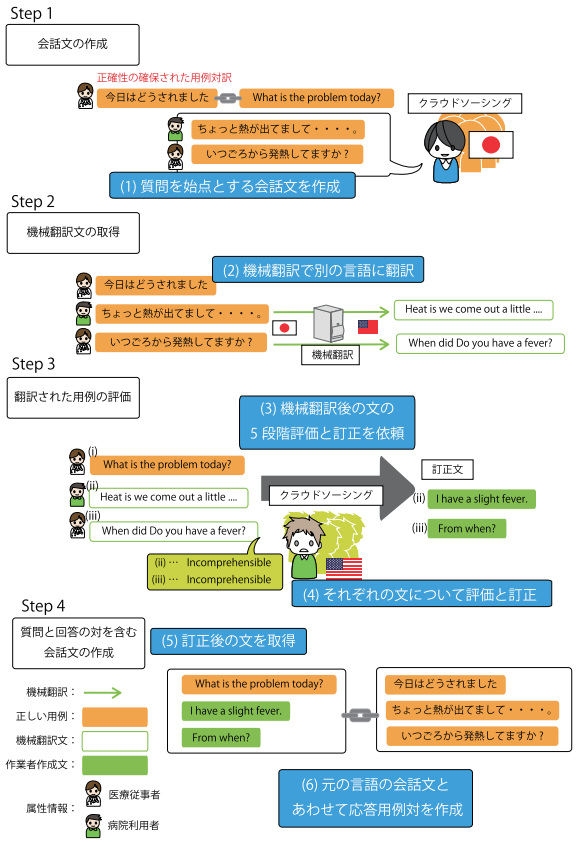

図 2 クラウドソーシングを用いた応答用例対作成の流れ

Fig. 2 The flow of creating a pair of dialogic parallel text using crowdsourcing.

Step 3 翻訳された用例の評価

図 2 の (3)，(4) で，クラウドソーシングによって機械翻訳文の評価，訂正文の入力を行うタスクを依頼する．図 2 の (4) で，作業者は提示された会話文の 2 文目と 3 文目に対して，流暢性評価を 5 段階で行い，訂正がある場合には訂正文を入力する．なお，流暢性評価の基準は，文献 [23] の評価基準[*5]を参考にした．文献 [23] の評価基準は文法に関して定義されている．しかし，本手法における流暢性評価は，文献 [23] とは異なり，評価対象文の文法だけではなく，会話文の文脈も考慮している．なお，評価の際には複数選択を可とする．

Step 4 質問と回答の対を含む会話文の作成

図 2 の (5)，(6) で，Step 3 で取得した評価と訂正文より，Step 1 で作成した文の用例対訳を作成し，それを質問と回答の対を含む会話文とする．

4. クラウドソーシングを用いた応答用例対作成実験

本章では，3 章で述べた手法の実験について述べる．

[*5] 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）

表 1 応答用例対作成実験に用いる 4 つの手法

Table 1 Four methods used for experiment to creating a pair of dialogic parallel text.

	会話文を提示	評価対象文のみ提示
使用者の属性情報あり	提案手法 A	従来手法 A
使用者の属性情報なし	提案手法 B	従来手法 B

4.1 応答用例対作成実験の概要

実験では，4.2 節で述べる実験用データセットの作成に，Yahoo!クラウドソーシング[*6]を，4.3 節で述べる応答用例対作成に CrowdFlower[*7]を使用した．Yahoo!クラウドソーシングは Yahoo!JAPAN の運営するクラウドソーシングサービスであり，作業者を Yahoo!クラウドソーシングサービスに登録した Yahoo!JAPAN ユーザとしている．また，作業者に支払われる報酬は T ポイント[*8]として支払われる．また，CrowdFlower は業務を細分化して作成した，数分～数十分で完了できるような単純作業のタスクを依頼するマイクロタスク型のクラウドソーシングサービスである．なお，機械翻訳サービスとして Google 翻訳[*9]を使用し，日本語から英語に翻訳した．また，本実験では，提案手法の有用性の確認のため，**表 1** に示す 4 つの手法において応答用例対作成実験を行った．文献 [22] では，作業者に用例の使用者の属性情報を提示することで，提示しない場合よりも正確に翻訳文の訂正が行える場合があることを示した．そのため，この 4 つの手法の比較を行うことによって，用例の使用者の属性情報ではなく，作業者に会話文を提示することの有用性の評価を行う．なお，本論文では，用例の使用者の属性情報として「医療従事者」と「病院利用者」を用いた．4 つの手法は以下，それぞれ提案手法 A，提案手法 B，従来手法 A，従来手法 B と呼ぶ．

4.2 実験用データセット作成のための会話文作成タスク

本節では，実験用データセット作成のために行った，Yahoo!クラウドソーシングにおける会話文作成タスクについて述べる．会話文作成タスクでは，以下の条件にあてはまる用例を会話の始点とした会話文の作成を依頼した．

- TackPad に登録されている日本語の用例である．
- 正確性の確保された英語の用例対訳がすでに登録されている．
- 英語の用例対訳に “What” または “How” を含む疑問文である．
- 英語の用例対訳が “How much～？” や “How many～？” などの，数値が回答となる疑問文ではない．

また，作業者に対して，評価を行う用例の使用者の属性

[*6] http://crowdsourcing.yahoo.co.jp/
[*7] http://crowdflower.com/
[*8] カルチュア・コンビニエンス・クラブが展開するポイントサービスであり，ポイントは商品や現金と交換することが可能．
[*9] https://translate.google.com/

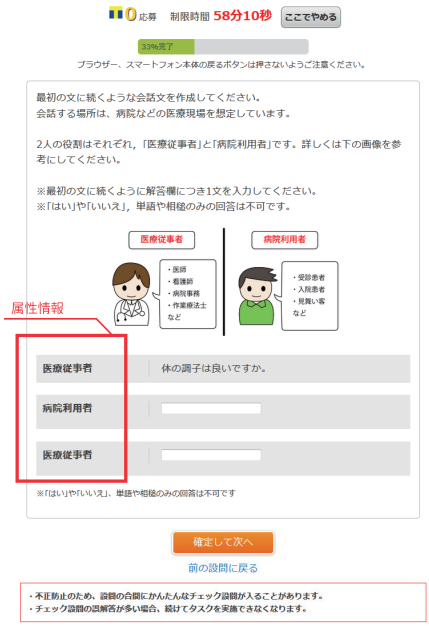

図 3 クラウドソーシングにおける質問と回答の会話文作成タスク画面

Fig. 3 Screenshot of creating a combination of question and answer using crowdsourcing.

情報を示した．この属性情報は，用例対訳として登録されている用例の主語や文脈などから，著者の 1 人が判断して決定したものである．用例の使用者の属性情報の判断は，それぞれの用例を 2 種類に分類するタスクのため，比較的容易に行うことができたが，まれに判断に迷うような用例も存在した．その場合は，用例対訳の主語などを見ることで使用者の属性情報を判断することができ，判断は機械的に行える作業であった．タスク画面の例を**図 3** に示す．作業者には会話の始点となる用例と，その用例が医療現場で使用されることを想定していることを提示し，作業者が作成する用例の使用者の属性情報を指定する．

条件にあてはまった 20 文について，それぞれを会話の始点として，続く文を 2 文追加する設問を作成した．1 タスクあたり 2 問の設問を設置し，2 組の会話文を作成するように設定した．また，1 文あたり 10 人の作業者に依頼した．1 タスクあたりの報酬は 5 ポイントとした．その結果，本タスクに対する請求金額は 1,512 円（税込み）[*10] となった．そのため，1 つの会話文作成あたり 7.56 円のコストがかかったことになる．また，会話文作成にあたって，作業者には以下の指示を行った．

- 医療現場で想定される会話文を作成すること
- 「はい」や「いいえ」のみの回答の禁止
- 相槌や，それのみでは回答として成立しない単語のみの回答の禁止
- 1 つの解答欄に 2 文以上記入することの禁止

[*10] 本実験では，Yahoo!クラウドソーシングの定める規定により，1 タスクあたり 14 円となった．

本実験では 200 組の会話文が取得できた．よって，作業者が作成した文は 400 文である．評価の際に区別するために，この 400 文に対して，1〜400 までの ID を付与した．取得した 400 文に対して，作成した文どうしや，TackPad に登録されている用例との類似判定を行った．TackPad に登録されている用例との類似度の判定は N-gram に基づく用例対訳検索手法を利用し [24]，類似文を類似度の高い順に最大で 3 文抽出する．文献 [24] では，用例を 2-gram もしくは 4-gram の文字に分割し，検索文字列との共起を調べることで，多言語の類似文検索を実現している．また，TackPad から抽出した用例と作成した文での類似度判定，作成した文どうしでの類似度判定には PHP の similar_text 関数を用い，類似度 80%以上のものを類似文とした．また，抽出した類似文を同一のものとし，TackPad にすでに登録されている場合は TackPad の用例を，それ以外の場合には作成されたのが 1 番早かった文を代表の 1 文とした．類似文判定で 80%以上の類似度が見られた文を同一のものとしたときの代表の文を実験に使用する．また，作成した文のうち，TackPad に登録されている用例と置き換えが可能なものは，TackPad の用例と置き換えて使用した．本実験では，取得した会話文の数が多く，すべての会話文について評価を行った場合，評価数が膨大なものとなることを防ぐため，類似文の統合を行っている．類似文の代表として TackPad に登録されている用例を使用した場合，その用例に用例対訳が存在していれば，作業者に提示する機械翻訳文の代わりに，すでに正確性の確保されている用例対訳を使用することができる．これにより，作業者に提示される会話文は，文法的に正しい文を 2 文以上含むこととなり，作業者が会話文の使用される状況をより推測しやすくなると考えた．実験に使用することにした，類似文を同一の文としたときの代表の文を用いて，クラウド作業者が作成した会話文に直したところ，38 組の会話文が取得できた[*11]．そのため，本実験で使用した会話文は 38 組である．

4.3 応答用例対作成実験

本節では，クラウドソーシングを用いた翻訳文の評価と訂正を行うことによって作成される評価データの収集について述べる．本実験では，提案手法 A，B においては，機械翻訳や TackPad に登録された用例対訳を用いて作成した英語の会話文のうち，2 文目と 3 文目の評価・訂正を行うことを 1 つの設問とし，1 タスクあたり 3 つの設問を作業者に依頼した．また，会話文を提示しない従来手法 A，B では，提案手法 A，B での 1 タスクあたりの評価文数と一致させるため，1 タスクあたり 6 文の評価を行う．作業

[*11] 類似文判定は 1 文ずつで行ったが，実験に使用するのは会話文のため，作業者によって作成された会話文の形に戻す必要がある．実験に使用することにした文のうち，1 組の会話文の形に戻せなかった文は今回の実験では使用しないため，最終的に実験に使用する会話文は 38 組の会話文となった．

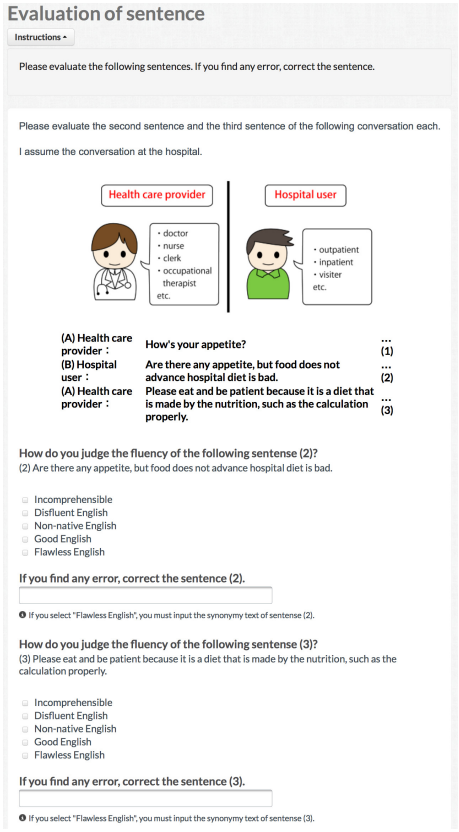

図 4 クラウドソーシングにおける応答用例対作成タスク画面
Fig. 4 Screenshot of creating a pair of dialogic parallel text using crowdsourcing.

者による英文評価には，機械翻訳によって作成された文を用いる．例外として，機械翻訳前の用例に，すでに用例対訳が存在している場合は，その用例対訳を用いた会話文を作成し，評価の依頼を行う．以降，作業者が評価を行う文を「翻訳文」とする．また，すべての手法において，1 つの翻訳文あたり 10 回評価が行われるように設定した．なお．4.2 節において取得した会話文は 38 組のため，評価対象となる翻訳文は 76 文である．1 つのタスクに対する報酬は 10 セントとし，1 人の作業者あたり 6 文の翻訳文の評価を行うよう設定したため，設問の数は，提案手法では 380 件，従来手法では 760 件となった．これより，本実験で行う総タスク数は，それぞれの手法で 127 件となった．そのため，タスクの依頼に必要な金額はどの手法も$16.85 となった．今回の実験では，1 組の会話文あたりの評価には，$0.443（日本円で約 45 円[*12]）のコストがかかった．なお，評価は CrowdFlower によって定められたレベル 3[*13]の，国を United States とする作業者に依頼した．提案手法 A におけるタスクの画面例を**図 4** に示す．作業者には，評価対象の翻訳文の使用者が医療従事者，または病院利用者であることを示し，会話文を提示した．作業者は，提示された会話文の 2 文目，3 文目について，5 段階の流暢性評価と，訂正が必要な場合には訂正文の入力を行う．

[*12] 2014/6/24 現在，1 ドル = 101.91 円で計算．
[*13] CrowdFlower は，作業者群を 3 段階の能力評価基準で分けている．レベル 3 は最も上のレベルである．

表 2 各手法において取得した評価数と訂正文の数
Table 2 The number of judgments and correction sentences acquired by using each method.

	提案手法 A	提案手法 B	従来手法 A	従来手法 B
(1) 取得した評価の数	760	760	820	760
(2) 不正な評価の数	137	139	105	108
(3) 有用な評価の数	623	621	715	652
(4) 有用な訂正文の数	478	483	457	402
(5) (4) で完全一致する文を除いた訂正文の数	393	409	388	326

取得した評価データのうち，以下の条件にあてはまる評価データは，不正な評価として本実験の考察には用いなかった．

- 同じ作業者が別の翻訳文に対して同一の訂正文を記入した場合．
- 2 単語以下の訂正文で，“not good” や “none”，“jkljlkj” などの明らかに不適切な文．
- 機械翻訳文の意味を推測し，説明している文．
- “I have no idea to fix it.” など，コメントをしている文．
- 5 段階評価で 3 以下と評価したにもかかわらず，翻訳文をそのまま使用している文．

5. 考察

本章では，4 章における実験の結果について考察を行う．

5.1 取得したデータの種類とその数，評価結果の例

本実験で取得したデータの種類とその数について，**表 2** に示す．表 2 の（2）は，前述の不正な評価の条件のいずれかにあてはまった評価であり，（3）は訂正文の数[*14]である．なお，従来手法のみ総評価数が 820 件となったのは，CrowdFlower の，作業者にランダムに unit[*15]を提示する仕様により，提示する unit に偏りが生じたことによると考えられる．なお，取得評価数は異なるが，タスクにかかったコストはすべて同じであった．また，本論文では，10 件以上の評価を得られた翻訳文[*16]については，そのまますべての評価を用いた．また，従来手法と提案手法を比較するために，会話文に含まれる 3 つの文を 1 つずつの文として

[*14] 提示した翻訳文と同じ文，または “Base sentence” と入力されていない訂正文．
[*15] CrowdFlower では各設問に用いるデータを 1 つの unit としている．unit の数は設問の種類の数である．
[*16] 本実験では，既存手法 A において，11 件以上の評価を得た翻訳文が 20 文あった．これは，利用しているクラウドソーシングサービスのシステムの仕様により，作業者に提示する会話文をランダムに提示するうえで，提示する翻訳文に偏りが生じ，11 回以上評価が行われたと考えられる．

表 3 翻訳文に対する評価結果の例
Table 3 Examples of the judgment result of the translated texts.

ID	評価対象文	提案手法 A		提案手法 B		従来手法 A		従来手法 B	
		最頻値	中央値	最頻値	中央値	最頻値	中央値	最頻値	中央値
13	It is not too much.	3	3	3	3	5	5	5	5
97	It is very concerned about attacks or not Shimawa happening at work.	2	2	2	2	2	2	2	2
167	First, is the diet.	2	3	2	2	4	4	5	5
173	Now my husband because towards us, you get paid when you get.	1	1	2	2	2	2	1	1.5
238	What did you eat yesterday?	5	5	5	5	5	5	5	5
266	When did my.	1	1	1	1	2	2	1	1

表 4 提案手法において作成された訂正文
Table 4 The correction sentences in the proposed methods.

元の文（機械翻訳文）	訂正文	
	提案手法 A	提案手法 B
It is not too much.	It’s not very good.	I’m not too hungry.
	It is not too great.	Is it not very good.
	I don’t have much of one.	Not too much.
	I do not have much of an appetite.	It is not too great.
	It’s lower than usual.	It’s not that big.
	It’s not much of an appetite.	I don’t have much of one.
		I don’t have much.
		It is not very good.

・元の文は，「余りありません。」に対して機械翻訳を行った結果である．
・元の文の日本語と対となる文は「食欲はありますか？」である．
・従来手法 A では訂正は行われなかった．
・従来手法 B では 2 文の訂正文が作成された．

扱い，2 文目と 3 文目の翻訳文の評価・訂正の結果をそれぞれ 1 つとして数える．

翻訳文に対する評価結果の例を**表 3** に示す．表 3 より，手法によって評価値に差がある場合と，差がない場合があることが分かった．5.2 節と 5.3 節で，それぞれの場合について述べる．

5.2 手法によって評価値に差がある翻訳文

本節では，表 3 の ID13 や ID167 のような，手法によって評価値に差が生じた翻訳文についての考察を行う．ID13 は，提案手法 A では評価が低いが，従来手法では評価が高い．これは “It is not too much.” が，文法的に誤りがなく，正しい機械翻訳が行われているためだと考えられる．しかし，提案手法では，ID13 に対応する疑問文として，作業者に “How’s your appetite?” が提示されていたため，それに対する回答としては不適切であると判断され，低い評価を行った可能性がある．**表 4** に，提案手法において作成された，ID13 の翻訳文に対する訂正文を示す．なお，本実験では従来手法 A で訂正文が作成されず，従来手法 B では “It will not be too much longer.” と “It is not too much” の 2 文のみ訂正文が作成された．従来手法では，ほとんど訂正は行われていない．

このことから，提案手法での作業者は，評価対象文の前の文に対する回答として適した文に訂正を行ったと考えられるため，提案手法による応答用例対の作成は有用であると考えられる．

5.3 手法によって評価値に差がない翻訳文

本節では，表 3 の ID97 や ID238 のような，手法によって評価値に差がない翻訳文について考察を行う．本実験において，提案手法 A と従来手法 A で評価が一致[*17]した評価の数は 28 文（38%）あった．どの手法でも評価値が高い翻訳文では，作成された訂正文の数は少ない．このため，機械翻訳が正しく行われた可能性がある．逆に，どの手法でも評価値が低く，訂正文の多い翻訳文では，機械翻訳が正しく行われていない可能性がある．たとえば，表 3 の ID97 では，元の文「仕事中に発作が起きてしまわないかがとても心配です。」に対し機械翻訳を行った結果「しまわ」の部分はローマ字表記にされたのみで，正しい翻訳が行われていない．また，ID97 では取得できた訂正文の数が多く，ID97 を評価した作業者の 80%が訂正文の入力を行った．取得した訂正文は 33 文であったが，誤った機械翻訳である “Shimawa” を使用していた訂正文は提案手法 A において 8 文中 1 文，提案手法 B，従来手法 A ではともに 3 文（提案手法 B で作成された訂正文は 9 文，従来手法 A で作成された訂正文は 8 文），従来手法 B では 8 文中 7 文であり，提案手法 A では 6 文が正しい訂正が行われていた．このことから，機械翻訳が正しく行われていない場合に，提案手法 A が有用である可能性があると考えられる．

表 5 で，評価値ごとの評価数に対する訂正文の割合の関係を示す．表 5 では，最頻値ごとに，取得した評価数に対する訂正文数を百分率で示している．表 5 より，どの手

[*17] 最頻値，中央値ともに同じ評価値となった評価データ

表 7 各手法における訂正文の違いの例
Table 7 Example of difference of correction sentences in each method.

ID	元の文（機械翻訳文）	訂正文		
		提案手法 A	提案手法 B	従来手法 B
13	It is not too much.	It's not very good.	**I'm not too hungry.**	It will not be too much longer.
		It is not too great.	Is it not very good.	It is not too much
		I don't have much of one.	Not too much.	
		I do not have much of an appetite.	It is not too great.	
		It's lower than usual.	It's not that big.	
		It's not much of an appetite.	I don't have much of one.	
		My appetite is not too great.	I don't have much.	
			It is not very good.	

・従来手法 A では，ID13 に対する訂正文は作成されなかった．
・応答用例対としてのみ有用な訂正文を太字で示す．

表 5 評価値ごとの評価数に対する訂正文の数の割合
Table 5 A ratio of the number of correction sentences to the number of judgments at each evaluation.

評価値（最頻値）	訂正文の数/評価数			
	提案手法 A	提案手法 B	従来手法 A	従来手法 B
5	55/157 (35%)	54/165 (33%)	63/276 (23%)	50/264 (19%)
4	46/99 (46%)	59/104 (57%)	46/94 (49%)	31/72 (43%)
3	91/127 (72%)	102/137 (74%)	78/103 (76%)	81/101 (80%)
2	128/165 (78%)	148/168 (88%)	176/213 (83%)	110/152 (72%)
1	69/75 (92%)	41/47 (87%)	22/29 (76%)	51/63 (81%)

・各項目の上段は取得した件数である．
・各項目の下段は割合（%）である．

表 6 訂正文の有用性の判断
Table 6 The judgment of usability of the correction sentences.

	提案手法 A		提案手法 B		従来手法 A		従来手法 B	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
作業者 A	344	317	354	323	322	310	268	261
作業者 B	357	348	366	356	343	348	280	281
作業者 C	251	236	256	249	217	210	174	182
平均	317	300	325	309	294	289	241	241
割合（%）	81	76	79	76	76	74	74	74

・評価を行った訂正文の総数は，提案手法 A では 393 文，提案手法 B では 409 文，従来手法 A では 388 文，従来手法 B では 326 文である．
・各手法の表中（1）の数値は，応答用例対として使用可能と判断された訂正文の数である．
・各手法の表中（2）の数値は，用例対訳として使用可能と判断された訂正文の数である．

法においても高い評価値を得られた翻訳文については，訂正文の数も少ない．そのため，応答用例対としても，用例対訳としても使用が可能であると考えられる．また，どの手法においても評価値の低い翻訳文においては，訂正文の数は多いが，正しく訂正を行えている訂正文の数が少ない場合がある．手法によって，評価数に対する訂正文の割合に大きな違いは見られないが，訂正の内容は，手法によって違いがあった．取得した訂正文の内容と有用性の考察は 5.4 節で行う．

5.4 各手法における訂正文の考察

本節では，作成された訂正文について，応答用例対として使用可能かどうかの判断と，用例対訳として使用可能かどうかの判定を行う．これは，文 ID13 において作成された，「余りありません。」の機械翻訳文 "It's not too much." に対する訂正文 "I do not have much of an appetite." のように，食欲に関する質問が行われた会話文の中の文としては意味が通じるため，応答用例対として使用可能だが，1 対 1 の用例対訳として使用可能ではない訂正文が作成されたためである．なお，評価は大学生 3 人に依頼した．**表 6** に評価者ごとの訂正文の評価結果と，各手法における使用可能な訂正文の数の平均，作成された訂正文数に対する割合を示す．

表 6 の割合より，提案手法の方が，従来手法で訂正文を作成するよりも多くの有用な訂正文を作成できると考えられる．また，提案手法 A および従来手法 A では，前処理として使用者の属性情報を付与したが，提案手法 B では使用者の属性情報は付与しなかった．しかし，同様に使用者の属性情報を付与していない従来手法 B よりも多くの訂正文の作成を行うことができた．そのため，提案手法を用いることで，使用者の属性情報の付与という手順を省いたとしても，有用な応答用例対や用例対訳が作成できると考えられる．これにより，応答用例対作成を行う際，本実験で行った用例の使用者の属性情報を判定するという，クラウドソーシング外での作業を減らすことができる可能性がある．しかし，応答用例対（会話文の中の 1 文として使用可能．ただし，元の日本語の意味としては成り立たない場合がある）として有用であっても，用例対訳（正確性の

表 8 各手法における英文作成にかかるコストの違い
Table 8 The difference in the cost creation English sentence in each method.

	総コスト（円）	作成文数（文）	1 文あたりのコスト（円）
(1) 提案手法 A	3,200	317	10
(2) 提案手法 B	3,200	325	10
(3) 既存手法 A	3,200	294	11
(4) 既存手法 B	3,200	241	13
(5) 専門家に依頼[*18]	–	–	925

・(1)〜(4) の総コストは，会話文作成と翻訳文評価にかかったコストの合計である．
・(1)〜(4) の作成文数は，表 6 の，応答用例対として有用と判断された訂正文数の平均を使用した．
・(5) のコストは文字数によるため，提案手法 A で作成された有用な応答用例対の平均文字数をもとにコストを算出した．

確保された多言語テキストペアであり，それぞれの文において，過不足なく翻訳が行われている）として有用ではない訂正文が作成された翻訳文が数件存在した．その例を，**表 7** に太字で示す．表 7 では，各手法における訂正文の違いの例として，ID13 に対して作成された訂正文を手法ごとに示す．

ID13 では，機械翻訳前の文は「余りありません。」であり，機械翻訳の前後にかかわらず，「食欲」や “appetite” などの，食に関する言葉は登場していない．しかし，ID13 の前にある文は “How’s your appetite?” であり，食に関する文である．このことから，作業者が，評価対象の翻訳文を訂正する際，会話文に適した文として書きかえようとしたためだと考えられる．なお，このような応答用例対にのみ使用可能な訂正文が作成された翻訳文は，ID13 以外にも ID266 など，全部で 11 文存在した．

よって，本手法を用いて作成した応答用例対の多くは，用例対訳としても使用することが可能であるが，なかには用例対訳としては機能しない文が存在する可能性があるため，本手法を用いて用例対訳の作成を行おうとする場合は注意が必要である．

6. クラウドソーシングのコスト

本章では，翻訳文の評価・訂正に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．表 6 の，応答用例対として使用可能と判断された訂正文数を，作成された英文として，コストの比較を行う．**表 8** に各手法と，専門家に依頼した場合のコストについて示す．専門家へ依頼した場合のコストの計算には，エキサイト翻訳依頼プロ[*18]で医療用例の翻訳を依頼した場合を想定してコストの算出を行う．エキサイト翻訳依頼プロでは，文の数ではなく文字数でコストを算出するため，提案手法 A における，使用可能な訂正文の平均文字数である 37 文字を，作成した英文の文字数とした．エキサイト翻訳依頼プロにおける，医療用例の翻訳には 1 文字あたり 25 円が必要である．このため，1 文の翻訳に約 925 円かかると考えられる．

CrowdFlower における応答用例対作成では，各手法ともに$16.85（日本円で約 1,700 円）のコストがかかった．本実験では，38 組の会話文を用い，76 文の評価を行ったため，1 文の翻訳文の評価・訂正には約 23 円がかかった．提案手法 A，B では 1 対の応答用例対の作成にかかったコストは約 10 円（それぞれの手法において約 320 文の応答用例対が作成できたため（表 6））であり，これは，翻訳者に依頼して翻訳を作成するよりも安価である．また，翻訳者による翻訳は対訳が 1 文のみ作成されるが，本手法では複数の対訳を作成することができる．複数の対訳を作成することができれば，自分の言葉の翻訳ではなく，相手のいったことを理解するための翻訳が行いやすくなるなど，自由度の高い利用ができる可能性がある．これらのことから，安価で多様性を持つ応答用例対や用例対訳の作成が可能な本手法は，有用であると考えられる．

[*18] https://orderpro.excite.co.jp/

7. おわりに

本論文では，クラウドソーシングを用いた応答用例対作成時に，作業者に対して，評価行う際に翻訳文を含む会話文を提示する手法の提案を行い，翻訳文評価の実験を行った．

本手法では，クラウドソーシングを用いた翻訳文の評価と訂正を行う際に，作業者に対して，評価対象の翻訳文を含む会話文を提示することで，正確な応答用例対の作成を行うことを目指した．今回の実験により，本手法が応答用例対作成を行える可能性があることが分かった．しかし，応答用例対として有用ではあるが，用例対訳としては使用できない訂正文がいくつかあった．そのため，本手法を用いて応答用例対を作成した場合，それを用例対訳として使用する際には注意が必要であることが分かった．

今後は，会話文を用いた方が訂正精度が上がる多言語テキストペアと，訂正精度が下がる多言語テキストペアの違いについて考察を行い，より最適なタスクの作成方法について検討する．

謝辞 本研究の一部は，JST A-STEP「多段クラウドソーシングを活用した多言語用例対訳プラットフォームの構築」，JSPS 科研費 24220002 および 26730105 の助成を受けた．

参考文献

[1] 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 2014-05-10）．
[2] 日本政府観光局（JNTO）：訪日外客数の動向（オンライン），入手先 ⟨http://www.jnto.go.jp/jpn/reference/tourism_data/visitor_trends/⟩（参照 2014-05-10）．

[3] 法務省：訪日外国人 2500 万人時代の出入国管理行政検討会議（オンライン），入手先 ⟨http://www.moj.go.jp/nyuukokukanri/kouhou/nyuukokukanri01_00103.html⟩（参照 2014-06-27）.
[4] 山田豊通，清水由美子，関 裕志：異言語間コミュニケーション支援システムの提案，武蔵野工業大学環境情報学部情報メディアセンタージャーナル，No.2, pp.40–45 (2001).
[5] Ishida, T.: Language Grid: An Infrastructure for Intercultural Collaboration, *IEEE/IPSJ Symposium on Applications and the Internet* (*SAINT-06*), pp.96–100 (2006).
[6] 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.22–33 (2012).
[7] 福島 拓，吉野 孝：正確かつ自由度を高めた多言語間対話支援を目的とした応答用例対構築モデル，情報処理学会論文誌，Vol.56, No.1, pp.219–227 (2015).
[8] 藤井薫和，重信智宏，吉野 孝：機械翻訳を用いた異文化間チャットコミュニケーションにおけるアノテーションの評価，情報処理学会論文誌，Vol.48, No.1, pp.63–71 (2007).
[9] Yamashita, N., Inaba, R., Kuzuoka, H. and Ishida, T.: Difficulties in Establishing Common Ground in Multiparty Groups using Machine Translation, *Proc. ACM Conference on Human Factors in Computing Systems* (*CHI'09*), pp.679–688 (2009).
[10] 宮田 玲，立見みどり，Anthony Hartley，影浦 峡，井佐原均：日英機械翻訳の精度改善と原文の読みやすさ向上のための日本語書き換えルールの作成と評価，言語処理学会，第 19 回年次大会発表論文集，pp.710–713 (2013).
[11] 井佐原均：クラウドソーシング後編集，*Japio YEAR BOOK*, pp.238–241 (2013).
[12] 宮部真衣，吉野 孝：翻訳不適切箇所指摘による翻訳リペア効率の改善効果の検証，情報処理学会論文誌，Vol.50, No.4, pp.1390–1398 (2009).
[13] 角田啓介，菱山玲子：コミュニケーション支援エージェントを組み込んだ多言語参加型ゲーミングシステムの設計，2010 年度人工知能学会全国大会，1C1-4 (2010).
[14] 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).
[15] 尾崎 俊，松延拓生，重野亜久里，吉野 孝：携帯端末を用いた多言語医療対話支援システムの開発，情報処理学会，第 73 回全国大会講演論文集，No.1, pp.215–217 (2011).
[16] 鹿島久嗣，馬場雪乃：ヒューマンコンピュテーション概説，人工知能，Vol.29, No.1, pp.4–11 (2014).
[17] Negri, M. and Mehdad, Y.: Creating a Bi-lingual Entailment Corpus through Translations with Mechanical Turk: $100 for a 10-day Rush, *Proc. NAACL HLT 2010*, pp.212–216 (2010).
[18] Callison-Burch, C.: Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk, *Proc. EMNLP 2009*, pp.286–295 (2009).
[19] Ambati, V. and Vogel, S.: Can Crowds Build Parallel Corpora for Machine Translation Systems?, *CSLDAMT '10, Proc. NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp.62–65 (2010).
[20] Resnik, P., Buzec, O., Kronrod, Y., Hu, C., Quinn, A.J. and Bederson, B.B.: Using Targeted Paraphrasing and Monolingual Crowdsourcing to Improve Translation, *Transactions on Intelligent Systems and Technology* (*TIST*), Vol.4, No.3, Article No.38 (2013).
[21] 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).
[22] 山本里美，福島 拓，吉野 孝：クラウドソーシング上における使用者の属性情報を用いた用例対訳生成手法の提案，電子情報通信学会技術報告，AI2013-37, pp.7–12 (2014).
[23] Walker, K., Bamba, M., Miller, D., Ma, X., Cieri, C. and Doddington, G.: Multiple-Translation Arabic (MTA) Part 1, *Linguistic Data Consortium*, Philadelphia (2003).
[24] 田淵裕章，坂本 廣，北村泰彦：N-gram に基づく用例対訳検索手法，信学技報，人工知能と知識処理研究会，Vol.108, No.441, pp.43–48 (2009).

山本 里美 （学生会員）

1992 年生．現在，和歌山大学システム工学部デザイン情報学科在学中．多言語用例対訳収集に関する研究に従事．

福島 拓 （正会員）

1986 年生．2008 年和歌山大学システム工学部デザイン情報学科中退．2010 年同大学大学院システム工学研究科システム工学専攻博士前期課程修了．2013 年同専攻博士後期課程修了．博士（工学）．現在，静岡大学大学院工学研究科数理システム工学専攻助教．協調作業支援の研究に従事．

吉野 孝 （正会員）

1969 年生．1992 年鹿児島大学工学部卒業．1994 年同大学大学院工学研究科修士課程修了．博士（情報科学）．現在，和歌山大学システム工学部教授．CSCW，HCI の研究に従事．