

多言語用例対訳共有システム TackPad の 評価機能の実現と評価

福島 拓[†] 吉野 孝^{††} 重野 亜久里[‡]

[†] 和歌山大学大学院システム工学研究科 ^{††} 和歌山大学システム工学部

[‡] 特定非営利活動法人 多文化共生センターきょうと

現在, 在日外国人数は年々増加しており, 多言語によるコミュニケーションの機会は増加している. コミュニケーションを行う際, 言語の違いは大きな障壁となる. また, 医療の分野では医療従事者と患者間での正確性の高い意思の疎通が重要である. このため, 支援を行う際に正確性の確保が難しい機械翻訳を用いることは適切ではない. そこで我々は, 多言語用例対訳を共有する TackPad の開発を行っている. 今回, TackPad の用例対訳群の評価機能の実現とその評価を行った. 本稿では, 試用実験の結果から次の知見を得た. (1) 評価者と評価閲覧者間での意思の疎通を行うため, 評価内容を明確にする必要がある. (2) 「おすすめ度」という言葉を用いると, 評価されることに対する抵抗感が減る可能性がある.

Evaluation Function of Multilingual Parallel-text Sharing System TackPad

Taku FUKUSHIMA[†] Takashi YOSHINO^{††} Aguri SHIGENO[‡]

[†] Graduate School of Systems Engineering, Wakayama University

^{††} Faculty of Systems Engineering, Wakayama University

[‡] Center for Multicultural Society Kyoto

Recently, the number of foreign residents in Japan is increasing. Consequently, the opportunity of communication among people whose native language are different increases. When people communicate with each other, the difference of the language is a huge barrier. Moreover, it is important to communicate reliably between medical workers and foreign patients in the medical field. Because it is difficult for machine translation to guarantee accuracy, it is not appropriate to use in the medical field. Therefore, we have developed a multilingual parallel-text sharing system, called TackPad. This paper presents the development and evaluation of the parallel-text evaluation function of TackPad. From the results of two experiments, we obtained the following findings. (1) To communicate smoothly between an evaluator and users, we have to specify an evaluation index. (2) The evaluation expression “Recommendation” can relax the uncomfortable by evaluation for evaluator.

1 はじめに

現在, 在日外国人数や訪日外国人数は年々増加している¹⁾. このため, 機械翻訳などの言語資源を組み合わせて利用できる仕組みである言語グリッドの活動が広がるなど^{2), 3)}, 言語の壁を越える活動が活発化している. 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⁴⁾. 日本語を理解できない外国人と日本人とのコミュニケーションは十分に行うことができないことが多いと考えられる.

日本語を理解できないことの影響が顕著に現れる分野の1つに医療がある. 医療分野では, わずかなコミュニケーション不足で医療ミスが発生する恐れがある. 日本語が通じない外国人と日本人

の医療従事者間でのやり取りは, 意思の疎通が十分に行えずに医療ミスが発生する可能性が高くなると考えられる. 現在, 日本語を理解できない外国人の支援は医療通訳者が行っている. しかし, 医療通訳者は慢性的な人員不足を抱えている. また, 通訳者の身分保障や通訳者自身のメンタルケアなどの問題が存在している⁵⁾.

情報技術を利用した医療分野の支援として, “日本語でケアナビ”^{*1}や多言語医療受付支援システム M³ (エムキューブ)⁶⁾がある. これらは, 正確な用例対訳を使用して医療分野の支援を行っている. 用例対訳とは, 用例を多言語に翻訳した多言語コーパスのことを指す. 本稿では, 多言語に

^{*1} <http://nihongodecarenavi.jp/>

翻訳された同じ意味の用例群を用例対訳群と呼ぶ。

我々は Web 上での多言語用例対訳の収集、共有、提供を目的とする多言語用例対訳共有システム TackPad(タックパッド)の開発を行い、試用実験で有用性を確認した⁷⁾。しかし、開発した TackPad が収集した用例対訳群の正確性は、用例作成者に依存している。このため、用例対訳群作成者以外による正確性の確保が必要であると考えられる。

そこで本稿では、TackPad に登録された用例を評価する機能の設計、構築を行う。また、構築した評価機能を利用して試用実験を行い、評価機能についての評価を行う。

2 関連研究

円滑な異文化間コミュニケーション支援を目指して、機械翻訳を用いた支援技術の研究が行われている⁸⁾。しかし、機械翻訳は正確性が必要な医療分野で利用可能な精度には達していない⁹⁾。

そこで現在、用例対訳による支援が行われている。用例対訳を利用したシステムとして、“日本語でケアナビ”がある。“日本語でケアナビ”は、介護に関する日本語と英語の用例対訳群を約 8000 個提供している。しかし、これらの用例対訳群の作成には多くの時間がかかっている¹⁰⁾。

そこで我々は用例対訳群の収集、共有、提供を目的とした用例対訳共有システム TackPad の開発を行った⁷⁾。しかし、TackPad には用例対訳群の正確性の確保に必要な評価機能が存在していない。

Web 上のシステムの評価機能の例として、Amazon^{*2}や価格.com^{*3}などがある。これらのサイトでは、販売している商品の評価や感想を、その商品を購入、所有している利用者が行っている。実際に商品を購入、所有している評価や感想は、その商品の購入を思案している別の利用者にとって非常に説得力のある情報であり¹¹⁾、評価の有用性を判別する研究も行われている¹²⁾。

しかし、本システムで評価を行うべき項目は単語や文であり、商品の評価とは以下のような違いがあると考えられる。このため、商品の評価と用例の評価では必要とする要件が異なる可能性がある。

1. 理解可能な言語の用例はすべて評価が可能のため、商品の評価よりも評価数が多くなる
2. 用例は場面や相手によって変化するため、判断基準が多くあり評価が難しい

*2 <http://www.amazon.co.jp/>

*3 <http://kakaku.com/>



図 1 TackPad の画面例

3. 評価対象が利用者が作成した用例であり、利用者が互いに評価を行うためトラブルが発生する可能性がある

そこで本研究では、登録された用例を利用者が評価する機能の構築を行い、以下の検証を行う。

1. 評価基準を評価者が決めることで、評価閲覧者との意思の疎通ができるか
2. 利用者が作成した用例に対して評価されることに対して抵抗感があるか

これらの検証により、多言語用例対訳群の評価に必要な要件を明らかにする。

3 システム設計

本章では、用例対訳共有システム TackPad の設計について述べる。次節以降、既に開発を行っているシステムの概要、今回開発を行った評価機能の設計の順に述べる。

3.1 システム概要

3.1.1 本システムの特徴

本システムは、多言語用例対訳を収集するため、画面インタフェースを多言語としている。収集言語は、日本語、英語、中国語、韓国語、ポルトガル語、スペイン語、ベトナム語、タイ語、インドネシア語の 9 言語である。また、PHP と MySQL を使用して Web 上で収集を可能としている。本システムの画面例を図 1 に示す。本システムは、用例収集をコミュニティで行うため、システムデザインにおいても遊び心を持たせて事務作業のような雰囲気をなくすようにしている。

3.1.2 主要機能

本システムの主要機能と利用者との関係を含めたシステム構成を図 2 に示す。本システムの主要機能は以下に示す 3 つである。

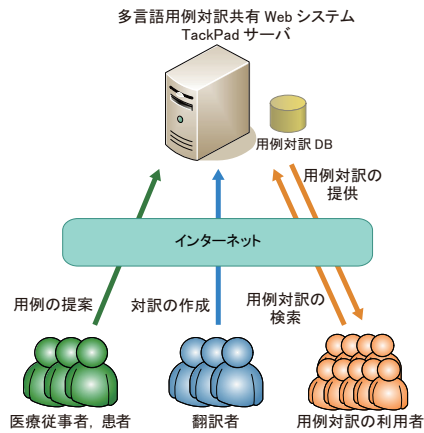


図 2 TackPad のシステム構成

おすすめ度

この文例のおすすめ度を、星をクリックして選んでください。



みんなのおすすめ度

5	5
4	1 3
3	2
2	1
1	0

図 3 評価機能 (おすすめ度) の例

1. 用例の提案

医療従事者や患者などが他の言語に翻訳してほしい用例を提案する機能である。実際に用例を使用する利用者がそれぞれの立場から提案するため、必要な用例を集めることができる。また、本機能は翻訳作業が不要なため、理解できる言語が1言語の利用者も本システムが目指している用例対訳群の収集、共有に貢献することが可能である。

2. 対訳の作成

用例の提案で提案された用例を翻訳者が翻訳する機能である。既に述べたとおり、医療分野では正確な翻訳が必要なため、本システムでは翻訳者が翻訳を行う。

3. 用例対訳の検索

本システム内の用例対訳群を検索する機能である。本機能は、医療従事者や患者、翻訳者などすべての利用者が利用可能である。

3.2 評価機能

評価項目は、(1) 用例の評価、(2) 用例間の評価の2種類を用意している。(1) の用例の評価は、用例の言い回しや誤字などを評価するために用意した。また、(2) の用例間の評価は、用例間の関係が

改善をお願いする文例

言語 日本語
 文例 採決室はどこですか？

修正依頼内容

依頼する内容を下から選んでください。

表記の変更依頼

- スペルが間違っています
- 文法が間違っています
- 言語選択が間違っています ?
- 入力文字に問題があります ?
- アクセント記号がありません

改訳依頼

- 親用例と異なる翻訳です
- 正確性に問題があります
- ニュアンスが違います

図 4 修正依頼機能の例

対訳として正しいを評価するために用意した。

本システムの評価機能の例を図3に示す。なお、本システムでは5段階評価で評価を行う。また、本システムでは評価機能を「おすすめ度」と表記している。これは、本システムの利用者は計算機に不慣れな人も多いと考えられた。このため、言葉を親しみやすくするために行っている。

実装した評価機能は星の上にマウスを乗せて動かすとリアルタイムに星の数が変化し、視覚的に星の数を確認可能である。また、クリックすることで評価を確定できるように設計した。なお、再度クリックすることで評価値の変更を可能とした。

また、利用者のうち正確な対訳を作成できると管理者が認定した利用者の評価を、他の利用者とは別に表示を行い、利用者ごとに評価の重み付けを可能とした。図3のみんなのおすすめ度の中で色が濃いものが認定された利用者の評価である。

なお、自由に記述できるコメント領域の追加も検討したが、利用者同士が評価し合うという本システムの特性上、否定的、攻撃的なコメントによる利用者間のトラブルが考えられたため実装を見送っている。

3.3 修正依頼機能

本システムに登録される用例は、利用者が登録を行う。しかし、言語の選択間違いやスペルミス、表現の不備が含まれる用例が登録される可能性がある。このため、登録された用例の修正が必要となるが、本システムでは管理者のみ他の利用者が作成した用例の編集を可能とし、一般利用者には

表 1 G2 の被験者の所属

被験者の所属	人数
医療関係者	1
通訳ボランティア	2
その他	2
無回答	1

編集権限を与えていない。

そこで、評価機能に付随する機能として、他人が作成した用例の修正依頼機能の構築を行った。修正依頼機能の例を図 4 に示す。本機能はチェックボックスで依頼を可能としているが、チェックボックスのみでは伝えきれない内容も存在するため、図 4 の次のページで表示される確認画面で自由記述を可能とした。ただし、前節でも述べたとおり、自由記述可能な領域は否定的、攻撃的な内容が記述される可能性がある。本機能では、送信者名が表示されるため修正依頼を送信した人を特定できる。また、本システムの利用者となるには管理者による事前の承認が必要であるため、現時点では大きな問題は生じないと考えているが、本機能については今後考察すべきであると考えられる。

4 試用実験

3 章で述べた正確性の確保を目的とした各機能の評価を行うために TackPad を用いて実験を行った。本実験の目的は、各機能の有用性の確認である。

4.1 被験者

今回行った実験は、以下に示す 2 種類のグループに所属する被験者が行った。

グループ 1(G1)：筆者らが所属する研究室の学生 10 人による評価

グループ 2(G2)：特定非営利活動法人多文化共生センターきょうとで募集を行った 6 人による評価 (想定される実際の利用者)

本稿では、グループ 1 に所属する被験者を G1、グループ 2 に所属する被験者を G2 とする。なお、G2 は、医療従事者や通訳ボランティアなどで、すべての被験者が複数言語話者である。G2 の被験者の所属を表 1 に示す。また、両グループの被験者は自宅に計算機を持っており、かつ、普段からインターネットを使用している人々である。

4.2 実験内容

本実験では本システムの想定利用方法を用意し、被験者には以下の順序で作業してもらっている。

1. 用例の提案 (新規登録) 作業

2. 用例への対訳登録作業 (翻訳者のみ)
3. 用例を検索し、検索結果の用例への評価付与
4. アンケート記入

アンケートの質問項目を表 2 に示す。なお、G2 の被験者のうち 1 人からは 5 段階評価のアンケートを得ることができなかったため、G2 の 5 段階評価については 5 名分の評価で行っている。

また、G1 の被験者については、自身が患者として病院に行った時の病院関係者との会話の場面を想定して用例の登録作業を行っている。

なお、評価者の属性による評価の重み付けは、本実験に参加する被験者の翻訳の質を事前に判断する必要があるので、本実験では行っていない。

5 実験結果の考察

本章では、試用実験の結果とその考察を行う。なお、各アンケート結果の後ろの括弧書きは、その内容を記述した被験者が属する被験者グループ名である。

5.1 評価機能

本節では、評価基準、評価に対する抵抗感、用例に付与された評価間の差の各項に分けて考察を行う。ただし、今回の実験に参加した被験者は全員が計算機を所持しており、計算機に関する知識が高いと考えられる。このため、計算機に不慣れな利用者に対しても同様の結果が得られるか今後確認する必要があると考えられる。

5.1.1 評価基準

評価機能の評価基準について、「どんな基準でおすすめ度をつけるのかが分からなかった (G1)」「何に対しておすすめ度をつけているかわかりにくい (G2)」などの意見を質問 1 の自由記述から得た。

設計時には自由に評価を可能にするために曖昧さを残していたが、これらの結果から、利用者にとっては曖昧さのせいで評価を行いにいく、また、評価された内容を閲覧する時も参考にしにくいことが分かった。このため、用例の評価を行う時は、評価内容を明確にして行う必要があると考えられる。

ただし、評価機能そのものについて、「使い方はわかりやすかった (G1)」「おすすめ度をつけることは簡単だった (G2)」などの意見を質問 1 の自由記述から得た。また、本システムには既に 5700 件を超える用例が登録されており、評価を多く行う必要がある。このため、機能は単純かつ簡単にする必要があると考えられる。このことから、1 回

表 2 アンケートの質問項目

質問	質問内容
1	用例を見て、5段階のうちどの“おすすめ度”にする(用例の質を判断する)か決めることは難しかった。
2	他の人が作った用例を閲覧する時に、“おすすめ度”を参考にした。もしくは、参考にと思う。
3	自分の作った用例に対して、“おすすめ度”が付くことに抵抗がある。
4	今後も用例に“おすすめ度”を付ける作業を行ってもいい。
5	“おすすめ度”の良かった点、改善点、問題点などをお願いします。
6	翻訳が間違えていたり言葉がおかしかったりした用例がTackPadに登録されていましたか？
7	間違いを含む用例を見つけた時、あなたはどうしましたか？

- ・質問 1~4 はリッカートスケールと自由記述、質問 5 は自由記述を用意した。
- ・質問 6 は選択項目として「はい」と「いいえ」を用意した。
- ・質問 7 は、「おすすめ度」を低くした、「修正依頼機能を使って用例の作者に報告した」「何もなかった」「その他」の選択項目と自由記述を用意した。また、質問 7 は質問 6 で「はい」と答えた人のみ答えるように誘導している。

のクリックで評価が完了する本機能の動作については有用であったと考えられる。

5.1.2 評価に対する抵抗感

質問 3(作成した用例におすすめ度が付くことに対する抵抗感)の結果を表 3 に示す。この質問では値が低いほど抵抗感が無いことを示している。また、G1 の被験者からは「(おすすめ度が) ついた方が嬉しいと感じる」「悪評でも自分のためになる」という意見を質問 3 の自由記述から得た。この結果から、G1 においては作成した用例に対しておすすめ度を付けられることについてはあまり抵抗がないことが分かる。これは、「おすすめ」という良い意味の言葉を使用したことや、「おすすめ」という言葉自体に「評価」という意味を感じさせない曖昧性があるため、抵抗感が無くなった可能性がある。このため、「おすすめ度」という言葉の持つ意味や曖昧性については今後検証する必要があると考えられる。また、G1 の被験者は対訳作成は行っていないため、対訳作成者の評価に対する抵抗感についても今後確認すべきであると考えられる。

なお、G2 の平均が G1 に対して高いが、「何を理由にしてその評価にしたか分からない」「どんな人に評価されているか分からない」などの意見を G2 の被験者から質問 3 の自由記述で得ている。このことから、評価に対する抵抗感よりも、評価を行った評価者の属性や評価基準が何であったかを重要視していると考えられる。

また、G1 の被験者からも質問 5 の自由記述で「(評価を付けた用例に) なぜその評価を付けたかをコメントで付けられると良い」という意見が複数出されている。設計時には利用者間のトラブルによる問題を想定していたためコメントを入力できる機能は実装していなかったが、システム利用者はどうしてその評価をしたのかを重要視している

表 3 作成した用例におすすめが付くことに対する抵抗感

	評価値					平均
	1	2	3	4	5	
G1	4	1	4	-	-	2.0
G2	-	1	4	-	-	2.8

- ・質問内容は、「自分の作った用例に対して“おすすめ度”が付くことに抵抗がある。」である。
- ・また、表中の平均とは、1:強く同意しない、2:同意しない、3:どちらとも言えない、4:同意する、5:強く同意する、の 5 段階で被験者に評価してもらった平均を表している。
- ・また、評価値の 1, 2, 3, 4, 5 の列は、上記の 5 段階評価の内訳の人数を表している。

表 4 2 回評価された用例の評価間の差

評価間の差	個数
0	6
1	8
2	1

ことが分かった。このため、コメントを入力できる機能は何らかの形で用意すべきと考えられる。ただし、トラブルが発生する可能性は残っているため、コメントを入力できる機能の実装方法を検討する必要があると考えられる。

5.1.3 用例に付与された評価間の差

実験終了後、2 回評価がされていた用例が 15 個あった。その用例に付与された評価間の差を表 4 に示す。また、質問 2 の自由記述で「他の人の文例のおすすめ度を参考に(評価を)した」という意見が G1 の被験者から得られた。これらから、既に入力された評価が他の評価に影響を与えていた可能性がある。このため、評価の前にこれまでの評価結果を提示しない等、評価の独立性を確保する工夫が必要となる可能性がある。

表 5 修正の必要な用例の確認の有無と、その後の対応

修正の必要な用例の有無，その後の対応		G1	G2
修正が必要な用例が登録されていた	おすすめ度を低くした	1	1
	修正依頼機能を利用した	1	0
修正が必要な用例が登録されていなかった		8	5

表中の数字は人数を表す。

5.2 修正依頼機能

表 2 の質問 6，質問 7(修正が必要であると感じた用例が含まれていた場合の，被験者の対応)の結果を表 5 に示す。この結果から，3 人の被験者が修正が必要であると感じる用例を発見したことが分かる。また，このうち 2 名が評価機能の評価を低くし，残りの 1 名が修正依頼機能を用いて用例作成者に修正依頼を行ったことが分かる。

用例の正確性の向上を目指すためには，評価を下げることもその用例の作成者に修正依頼を行うことが重要であると考えられる。今回の実装では評価機能と修正依頼機能を独立して実装していたが，修正依頼を行うように両機能を連携させる必要があると考えられる。

5.3 今後の評価機能の方針

本節では，5.1 節，5.2 節の実験結果の考察から今後の評価機能の方針について述べる。

アンケートの質問 5 の自由記述に「この例は良いと思ったものに 1 ポイントずつ入れるといい(G2)」という意見があった。また，5.1 節で評価基準を明確に表示した方がいいという結論に至った。このため，現状の 5 段階評価ではなく「使えそう」「対訳がほしい」「用例がおかしい」「対訳が間違っている」などの評価項目を書いたボタンを用意し，その中で利用者にあてはまる項目をクリックしてもらうのが良いと考えられる。この方法では，1 回のクリックで評価できるという簡単さは残したまま，評価基準の明確化が可能であると考えられる。

また，利用者が「用例がおかしい」「対訳が間違っている」などの，用例として問題があるボタンをクリックした場合は修正依頼機能へ誘導することで，用例の精度向上につながると考えられる。

6 おわりに

本稿では，多言語用例対訳共有システム TackPad への評価機能の実現と，試用実験を行った。

今回の試用実験では，次の知見を得た。

1. 評価者と評価閲覧者間での意思の疎通を行うため，評価内容を明確にする必要がある。

2. 「おすすめ度」という言葉を用いると，評価されることに対する抵抗感が減る可能性がある。

今後，提案した新しい評価機能を構築し，再度評価実験を行う。

謝辞

実験に参加していただいた，NPO 多文化共生センターきょうとの皆様，和歌山大学吉野研究室の学生の皆様に深く感謝を表す。また，実験の実施に多大なるご協力をいただいた多文化共生センターきょうとの前田華奈氏に心より感謝申し上げる。

なお，本研究の一部は総務省の戦略的情報通信研究開発推進制度(SCOPE)の平成 20 年度採択課題「多言語共生社会における医療対話支援のための多言語対話用例プラットフォームの構築」による。

参考文献

- 1) 法務省: <http://www.moj.go.jp/>
- 2) Toru Ishida: Language Grid: An Infrastructure for Intercultural Collaboration, IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06), pp.96-100(2006).
- 3) Satoshi Sakai, et al: Language Grid Association: Action Research on Supporting the Multicultural Society, ICKS-08, pp.55-60(2008).
- 4) 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)。
- 5) 高嶋愛里: “在日外国人支援活動: 京都における「医療通訳システムモデル事業」”, 国際保健支援会 2 2005.1(2005)。
- 6) 宮部真衣ほか: 多言語医療受付支援システムの構築と医療機関への導入, 信学研報, AI2008-35, pp.65-70(2008)。
- 7) 福島拓ほか: 医療分野を対象とした多言語用例対訳収集 Web システム TackPad の開発, マルチメディア, 分散, 協調とモバイル(DICOMO2008)シンポジウム, pp.1030-1036(2008)。
- 8) Takashi Yoshino, et al: Availability of Web Information for Intercultural Communication, Pacific Rim International Conference on Artificial Intelligence (PRICAI-08), pp.923-932(2008).
- 9) 林田尚子ほか: 翻訳エージェントによる自己主導型リペア支援の性能予測, 信学論, Vol.J88-D1, No.9, pp.1459-1466(2005)。
- 10) 上田和子ほか: 『日本語でケアナビ』と実践的コミュニティー, 国際交流基金関西国際センター日本語教育シンポジウム, パネルディスカッション資料, 泉南郡田尻町(2008-03-08)。
- 11) 渡辺弘美: ウェブを変える 10 の破壊的トレンド, ソフトバンククリエイティブ(2007)。
- 12) 山澤美由起ほか: Amazon レビュー文の有用性判別実験, 情処研報, 2006-NL-173-(3), pp.15-20(2006)。