情報処理学会論文誌 Vol.53 No.8 2036–2043 (Aug. 2012)

推薦論文

折り返し翻訳を用いた高精度なコミュニケーションのための複数翻訳機による精度不一致検出サービスの提案

宮部 真衣[1,a)] 吉野 孝[2,b)]

受付日 2011年12月5日, 採録日 2012年5月12日

概要：機械翻訳を介したコミュニケーションでは，翻訳精度が低い場合，十分な相互理解ができない可能性が高い．現在，母語のみを用いて自分の発言がどのように伝わっているのかを把握するための手法として，折り返し翻訳が用いられている．対象言語翻訳文と折り返し翻訳文の精度の同等性に関する検証の結果，対象言語翻訳文が不正確であるにもかかわらず，折り返し翻訳文が正確であるという状況（精度不一致）が発生する場合があることが明らかになっている．このような精度不一致が発生した場合，ユーザの確認する折り返し翻訳文には問題がないため，対象言語翻訳文が低精度であることに気づくことができず，大きな問題となる．そこで本研究では，複数翻訳機を利用した精度不一致の検出手法について検討を行う．検証の結果，提案手法とあわせて対象言語翻訳文中に原言語表現が残っているかどうかを検証する処理を行うことにより，約71%の精度不一致を検出できることを示した．

キーワード：多言語間コミュニケーション，機械翻訳，折り返し翻訳，精度不一致検出

Integrated Evaluation Using Multiple Translation Systems to Detect Mismatches between Back-translated and Target-translated Sentences

Mai Miyabe[1,a)] Takashi Yoshino[2,b)]

Received: December 5, 2011, Accepted: May 12, 2012

Abstract: In communication using machine translations, inaccurate translations can lead to misunderstandings. Therefore, it is important to check the accuracy of translations. Back translation is used to verify the accuracy of a sentence translated to a native language. However, a mismatch of accuracy that a translated sentence is inaccurate but its back-translated sentence is accurate sometimes occurs. In this case, people do not understand that a translated sentence is inaccurate because its back-translated sentence is understandable. We found that this mismatch can lead to serious problems in communication. Therefore, we proposed a method for detecting the mismatch in order to prevent such problems. The method obtains multiple back-translated sentences from different translation systems, and judges the accuracy of the translated sentence in a comprehensive manner. We found that the method can detect approximately 71% of the mismatches if combined with prior processing which checks the presence of letters of a source language in the target-translated sentence.

Keywords: multilingual communication, machine translation, back translation, detection of accuracy mismatches

1. はじめに

世界規模のインターネットの普及により，ネットワーク

[1] 東京大学知の構造化センター
Center for Knowledge Structuring, The University of Tokyo, Bunkyo, Tokyo 113–8656, Japan
[2] 和歌山大学システム工学部
Faculty of Systems Engineering, Wakayama University, Wakayama 640–8510, Japan
a) miyabe@cks.u-tokyo.ac.jp
b) yoshino@sys.wakayama-u.ac.jp

本論文の内容は 2011 年 7 月のマルチメディア，分散，協調とモバイル（DICOMO2011）シンポジウム 2011 にて報告され，グループウェアとネットワークサービス研究会主査により情報処理学会論文誌ジャーナルへの掲載が推薦された論文である．

を介した多言語間コミュニケーションの需要が高まっている．しかし，一般に多言語を十分に習得することは容易ではない．母語以外の言語を用いて十分なコミュニケーションを行うことは困難であり，相互理解ができない可能性が高い [1], [2]．母語でのコミュニケーションを支援するために，機械翻訳を用いた支援が行われている [3], [4]．

機械翻訳技術は急速に進展してきているものの，高精度な翻訳を行うことは困難である．機械翻訳を介したコミュニケーションでは，翻訳精度が低い場合，十分な相互理解ができず，思い違いが発生する [5]．このような思い違いを回避するためには，自分の発言がどのように伝わっているのかを把握する必要がある．母語のみを用いた対象言語の翻訳精度の把握は，折り返し翻訳（対象言語翻訳結果の原言語への再翻訳）を利用することにより実現可能である．折り返し翻訳は，機械翻訳を介した多言語間コミュニケーション支援において，精度確認手法として用いられている [6], [7]．

これまでに，翻訳精度確認手法としての折り返し翻訳の妥当性の検証が行われている [8]．検証の結果，折り返し翻訳文と対象言語翻訳文の精度には正の相関があることが示されている．一方で，対象言語翻訳文が不正確であるにもかかわらず，折り返し翻訳文が正確であるという状況（第 1 種の精度不一致）が発生する場合があることも明らかになった．この不一致が発生した場合，ユーザの確認する折り返し翻訳文には問題がないため，対象言語翻訳文が低精度であることに気づくことができず，大きな問題を引き起こす可能性が高い．

そこで，本論文では，第 1 種の精度不一致を検出するための仕組みを提案し，提案手法の効果について述べる．

以下，2 章において折り返し翻訳の課題について述べる．3 章で提案手法について述べる．4 章で検証実験について述べ，5 章で実験結果を示す．6 章で実験結果についての考察を述べる．最後に 7 章で本論文の結論についてまとめる．

2. 折り返し翻訳とその課題

折り返し翻訳とは，対象言語へと翻訳した結果を，原言語へと再翻訳することである．折り返し翻訳の流れを**図 1** に示す．対象言語に関する知識がない場合でも，折り返し翻訳文を確認することにより，対象言語翻訳文の翻訳精度を確認することができる [8]．また，翻訳自動評価において，対象言語の参照訳を用意せずに，翻訳精度を算出することができる [9], [10]．しかし，原言語への再翻訳によって得られる折り返し翻訳文は，「原言語から対象言語への翻訳」および「対象言語から原言語への翻訳」という，2 回の翻訳を介している．そのため，「対象言語から原言語への翻訳」を行うことにより，対象言語翻訳文の意味と折り返し翻訳文の意味が同一でなくなる可能性がある．

我々はこれまでに，翻訳精度確認手法としての折り返し翻訳の妥当性の検証を行った [8]．この研究では，妥当性の検証にあたり，以下の 2 種類の精度不一致を定義した．

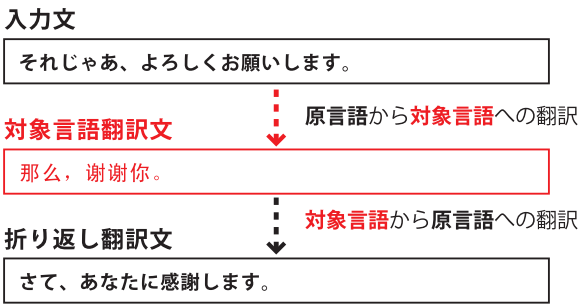


図 1 折り返し翻訳の流れ
Fig. 1 Procedure of back translation.

[第 1 種の精度不一致]：折り返し翻訳文の精度が 高い が，対象言語翻訳文の精度が 低い

[第 2 種の精度不一致]：折り返し翻訳文の精度が 低い が，対象言語翻訳文の精度が 高い

第 1 種の精度不一致が発生すると，入力者は伝わったと判断した内容が，相手の言語では正しく伝わらない．一方，第 2 種の精度不一致が発生すると，実際は修正しなくても伝わる文を，伝わらないと判断してしまう可能性がある．この場合，ユーザは本来不要な修正作業などを行う可能性があるが，第 1 種の精度不一致のような，意思疎通などの問題の発生にはつながらない．

検証の結果，第 1 種の精度不一致の発生率は低いものの，0%ではないことが分かった．第 1 種の精度不一致の発生は，意思疎通の阻害などを引き起こす可能性が高い．そのため，第 1 種の精度不一致が発生した場合の対策を講じる必要がある．

3. 第 1 種の精度不一致の検出手法

第 1 種の精度不一致では，ユーザの確認する折り返し翻訳文の精度が高いため，ユーザ自身が第 1 種の精度不一致の発生に気づくことは難しい．そのため，第 1 種の精度不一致を回避可能な翻訳サービスをユーザに提供することが望ましい．第 1 種の精度不一致を回避するためには，まず，第 1 種の精度不一致の発生を検出する必要がある．

そこで本論文では，第 1 種の精度不一致を検出するための手法を検討する．折り返し翻訳文と対象言語翻訳文の精度検証実験 [8] においては，原言語から対象言語への翻訳および対象言語から原言語への翻訳を行う際に，単一の翻訳システムを利用し，精度不一致の発生について検証を行った．しかし，第 1 種の精度不一致は，両方向（「原言語から対象言語」および「対象言語から原言語」）の翻訳において同じ手法が採用されている場合や，同じ言語資源から作られている場合などに発生しやすい可能性がある．たとえば，**表 1** に示す第 1 種の精度不一致の例では，原文中の「行なう」という表現の対訳として，韓国語翻訳文では

表 1 第 1 種の精度不一致の発生例
Table 1 Example of an accuracy mismatch between a target-translated sentence and its back-translated sentence.

原文	研究会は第五教室において 行なう。
韓国語翻訳文（システム A）	연구회는 다섯째 교실에서 지휘한다.
折り返し翻訳文（システム A）	研究会は、五番目の教室で 行なう。
折り返し翻訳文（システム B）	研究会は五つ目教室で 指揮する。
折り返し翻訳文（システム C）	研究会は五番目教室で 指揮する.

表中の対象言語翻訳文では，原文中の「行なう」という語が，「指揮する」を意味する表現（下線部）になっている．

「指揮する」を意味する語（表 1 の韓国語翻訳文における下線部）が用いられている．そのため，表 1 の韓国語翻訳文は文としておかしいと判定される．しかし，単一の翻訳システムを用いて折り返し翻訳文を生成すると，原文と同じ「行なう」という表現に戻っている．単一の翻訳システムを用いたことにより，「指揮する」を意味する語が，再び「行なう」に翻訳されたと考えられる．一方，対象言語翻訳時と異なる翻訳システムを用いた場合，折り返し翻訳文中に「指揮する」という表現が現れている．このように，単一のシステムによって得られた折り返し翻訳文の精度が高い場合に，複数の翻訳システムを用いて折り返し翻訳文を生成し，比較することによって，精度不一致の発生を検出できる可能性があるのではないかと考えた．

なお，複数の翻訳システムを利用する場合，複数の対象言語翻訳文を生成し，より高精度な対象言語翻訳文をユーザに提供することにより，対象言語翻訳文の精度が低いことで生じる第 1 種の精度不一致そのものを回避するという方法も考えられる．しかし，複数翻訳機によって生成した対象言語翻訳文の中に，翻訳精度の良い文があるとは限らない．生成した複数の対象言語翻訳文のいずれも翻訳精度が低い場合，第 1 種の精度不一致が発生する場合もあるため，それらに対して第 1 種の精度不一致が発生していないかどうか確認する必要がある．そのため，複数翻訳機を利用して，より高精度な対象言語翻訳文と折り返し翻訳文を提供するような仕組みを提供するとしても，第 1 種の精度不一致の検出は不可欠である．

そこで，複数翻訳機を用いた折り返し翻訳の精度不一致検出手法を提案する．提案手法を用いた折り返し翻訳提示の流れを**図 2** に示す．生成した折り返し翻訳文の翻訳精度が高精度である場合，複数翻訳機を用いて折り返し翻訳文の再生成を行う（図 2 手順 (4)）．それらの翻訳精度から総合的に精度を判定し，ユーザに提示する折り返し翻訳を選択することにより（図 2 手順 (5)），第 1 種の精度不一致の検出を目指す．

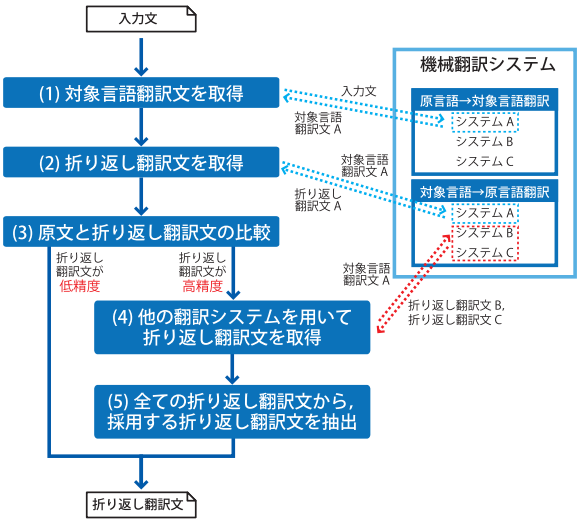


図 2 提案手法の流れ
Fig. 2 Procedure of our proposed method.

4. 検証実験

提案手法による第 1 種の精度不一致の検出効果を検証するために，実験を行った．

4.1 精度評価方法

折り返し翻訳文，対象言語翻訳文の主観評価は，Walker らの適合性評価（5 段階評価）[11] により行う．評価指標は，5：同じ意味，4：文法などに多少問題があるが，大体同じ意味，3：意味は何となくつかめる，2：雰囲気は残っているが，もとの意味は分からない，1：まったく違う意味，となっている．上記の評価基準を用いて，2 つの文（原文および折り返し翻訳文）の意味の比較を行う．評価者は，日本人大学生 3 名である．

なお，本論文では，上記の評価基準において，3 未満の場合は「意味が理解できない」，3 以上の場合は「意味が理解できる」と分類することとし，折り返し翻訳文の評価結果が 3 以上かつ対象言語翻訳文の評価結果が 3 未満である場合に第 1 種の精度不一致が発生していると見なす．

4.2 評価テキストおよび翻訳システム

Walker らの適合性評価による折り返し翻訳文の評価結果が **4 以上**である場合，折り返し翻訳文は原文の意味を持った文章になっている．そこで，折り返し翻訳の精度検証実験 [8] で用いられたテキストのうち，折り返し翻訳文の評価結果が「4」以上であったテキストを評価テキストとして用いる[*1]．また，折り返し翻訳文の評価結果が「5」の場合，原文と折り返し翻訳文の見た目にほとんど違いがない場合や，原文と折り返し翻訳文が同一である場合が多い．

[*1] 評価テキストの対象言語翻訳には，英語，中国語，韓国語が用いられている．また，折り返し翻訳文の生成には，対象言語翻訳時と同一のシステムを利用している．

表 2 評価テキスト数
Table 2 Number of sentences in evaluation texts.

	精度不一致文（文）	精度一致文（文）	合計（文）
評価テキスト A	19	276	295
評価テキスト B	64	379	443
合計	83	655	738

評価テキスト A：折り返し翻訳文の精度評価値が「5」のもの
評価テキスト B：折り返し翻訳文の精度評価値が「4 以上 5 未満」のもの

表 3 評価テキストの一部
Table 3 Examples of sentences used in the evaluation.

(1) 彼は駅からの距離を計った。
(2) 彼は私の顔をつぶした。
(3) とてもお財布にやさしいですね。
(4) チョコレート菓子ではなくてチョコレートですか？

このような場合，ユーザは第 1 種の精度不一致の発生に気づくことができないため，多言語コミュニケーションにおける問題を引き起こす可能性がきわめて高い．そこで，今回の検証においては，評価結果が「5」の文と「4 以上 5 未満」の文を分けて扱うこととする．以降，本論文では，折り返し翻訳文の評価結果が「5」の文を評価テキスト A，「4 以上 5 未満」の文を評価テキスト B と呼ぶ．評価テキストには，対象言語翻訳文の翻訳精度が低く，第 1 種の精度不一致が発生しているもの（精度不一致文）と，対象言語翻訳文の翻訳精度も高く，第 1 種の精度不一致が発生していないもの（精度一致文）が含まれている．

評価テキスト数を**表 2** に，評価テキストの一部を**表 3** に示す．評価テキストには，2 種類の文（機械翻訳試験文[*2]およびチャットにおける発言[*3]）が含まれている．

本実験では，3 種類の翻訳システム[*4,*5,*6]を用いて，評価テキストの折り返し翻訳文の生成（図 2 における手順 (4)）を行った．3 つのシステムのうち，2 つはルールベース翻訳システム，1 つは統計翻訳システムである．なお，各翻訳システムは言語グリッド [12] を介して利用した．

4.3 検証の流れ

検証の流れを**図 3** に示す．4.2 節で述べた評価テキスト（対象言語翻訳文および折り返し翻訳文）は，対象言語翻訳および折り返し翻訳時に同一の翻訳システムを用いて生成している．また，これらの対象言語翻訳文および折り返し翻訳文については，すでに精度評価がなされている．同一の翻訳システムを用いた場合の折り返し翻訳文の精度は 4 以上であり，対象言語翻訳文の精度に応じて精度不一致文（対象言語翻訳文の精度が 3 未満），精度一致文（対象言語翻訳文の精度が 3 以上）に分類されている．そこで，以下の手順により，精度不一致の検出・誤検出を検証する．

手順 1 ある対象言語翻訳文について，対象言語翻訳文生成時に用いたものとは異なるシステム（2 種類）によって折り返し翻訳文を生成する．

手順 2 手順 1 で生成した 2 つの折り返し翻訳文の精度を 4.1 節で述べた指標により評価する．

手順 3 3 つの折り返し翻訳文の精度評価値（手順 2 での 2 つ折り返し翻訳文の精度評価値と，対象言語翻訳時と同一のシステムによる折り返し翻訳文の精度評価値）の代表値から，精度不一致の検出・誤検出を検証する．

手順 3 では，3 つの折り返し翻訳文の精度評価値から，代表値を決める必要がある．代表値としては，中央値，最頻値，最小値，最大値などがある．今回は，最も検出効果が高くなる代表値として最小値を，データの中央にあたる値として中央値を用いる[*7]．

手順 3 においては，精度不一致文の場合，代表値が 3 未満であれば検出成功，3 以上であれば検出失敗と判定する．また，精度一致文の場合，代表値が 3 未満であれば誤検出，3 以上であれば誤検出なしと判定する．

5. 実験結果

5.1 検出率と誤検出率

中間値および最小値を代表値とした場合の検出率について検証する．

中央値を用いた場合の実験結果を**表 4** に，最小値を用いた場合の実験結果を**表 5** にそれぞれ示す．

表 5 より，最小値を用いた場合，評価テキスト A については 57.9%（19 文中 11 文），評価テキスト B については 65.6%（64 文中 42 文），全体では 63.9%（83 文中 53 文）の精度不一致を検出できた．表 4，表 5 より，最小値を代表値とした場合の検出率の方が，中央値を代表値とした場合よりも高い．一方，最小値を代表値とした場合，誤検出率も高くなった．

5.2 検出失敗の原因

表 4，表 5 に示したように，評価テキスト全体としての精度不一致の検出率は，中央値を用いた場合に約 35%，最小値を用いた場合に約 64%となっており，検出に失敗している文が存在する．そこで，精度不一致を検出できなかった精度不一致文に関して，対象言語翻訳文にどのような特

[*2] NTT Natural Language Research Group, http://www.kecl.ntt.co.jp/icl/mtg/resources/index.php
[*3] チャットにおける発言とは，「好きなもの・嫌いなもの」というテーマでのチャットにおける対話文である．
[*4] J-Server（高電社，http://www.kodensha.jp/）
[*5] Google 翻訳（Google，http://translate.google.co.jp/）
[*6] WEB-Transer（クロスランゲージ，http://www.crosslanguage.co.jp/）
[*7] 今回は 3 つの値の代表値を算出するため，最頻値が存在する場合，最頻値は中央値と一致する．そのため，最頻値ではなく中央値を用いる．

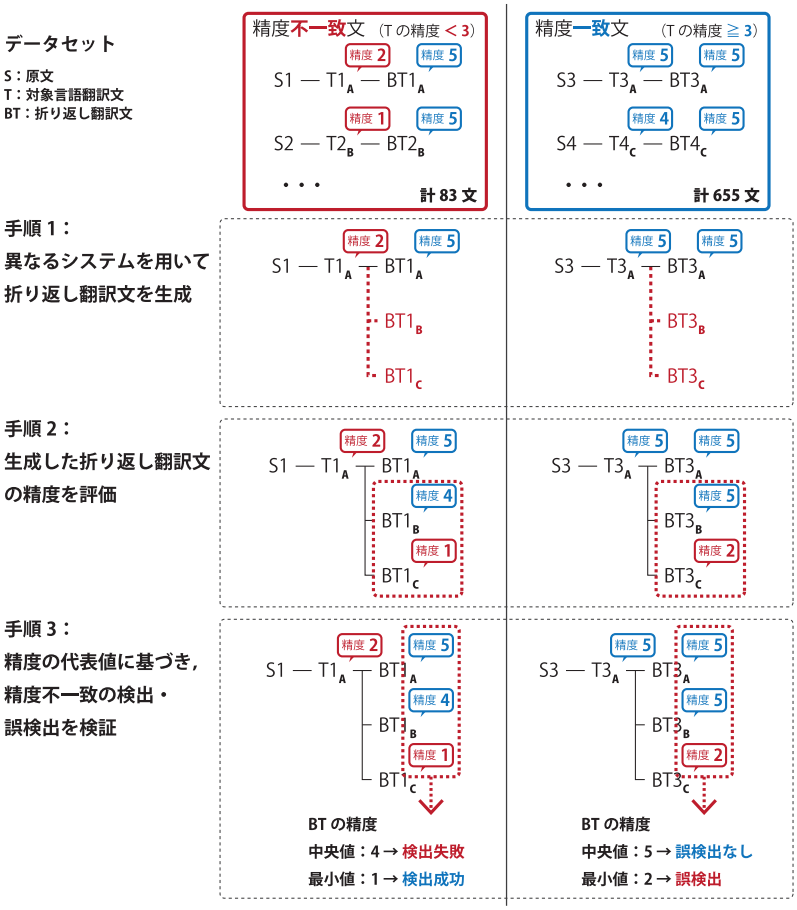


図 3 検証実験の流れ

Fig. 3 Procedure of experiment.

表 4 第 1 種の精度不一致の検出率および誤検出率（中央値を用いた場合）

Table 4 Detection rate and false-detection rate using median accuracy.

評価テキスト	折り返し翻訳文の精度評価値の中央値	精度不一致文（文）	精度一致文（文）	検出率（%）	誤検出率（%）
評価テキスト A	3 未満	5	15	26.3	5.4
	3 以上	14	261		
	合計	19	276		
評価テキスト B	3 未満	24	22	37.5	5.8
	3 以上	40	357		
	合計	64	379		
全体	3 未満	29	37	34.9	5.7
	3 以上	54	618		
	合計	83	655		

徴があるのかを確認した．対象言語翻訳文の確認については，各対象言語（英語，中国語，韓国語）の翻訳者および各対象言語を母語とする留学生に行ってもらった．

確認の結果，検出に失敗した対象言語翻訳文には，以下の傾向があることが分かった．

傾向 (A) 対象言語翻訳文が文として成立しない（原言語の表現が残っている）．

傾向 (B) 対象言語翻訳文が文として成立しない（語句の翻訳，文法に問題がある）．

傾向 (C) 対象言語翻訳文は，文として成立しているが，原文と意味が異なる．

各傾向に該当する検出失敗数を**表 6** に示す．

各傾向について，以下において説明する．

5.2.1 傾向 (A)

傾向 (A) による検出失敗例を**表 7** に示す．この例では，折り返し翻訳文は原文とまったく同じである．一方，対象言語翻訳文には原言語の表現が残っており，対象言語翻訳文の翻訳精度が低いと評価された．

実験においては，傾向 (A) による検出失敗数は，中央値を用いた場合は 54 文中 11 文（評価テキスト A が 5 文，評

表 5 第 1 種の精度不一致の検出率および誤検出率（最小値を用いた場合）
Table 5 Detection rate and false-detection rate using minimum accuracy.

評価テキスト	折り返し翻訳文の精度評価値の最小値	精度不一致文（文）	精度一致文（文）	検出率（%）	誤検出率（%）
評価テキスト A	3 未満	11	60	57.9	21.7
	3 以上	8	216		
	合計	19	276		
評価テキスト B	3 未満	42	108	65.6	28.5
	3 以上	22	271		
	合計	64	379		
全体	3 未満	53	168	63.9	25.7
	3 以上	30	487		
	合計	83	655		

表 6 検出に失敗した文の傾向
Table 6 Causes of detection failure and number of sentences of detection failure.

	中央値を用いた場合			最小値を用いた場合		
	評価テキスト A（文）	評価テキスト B（文）	全体（文）	評価テキスト A（文）	評価テキスト B（文）	全体（文）
傾向 (A)	5	6	11	4	2	6
傾向 (B)	6	28	34	2	18	20
傾向 (C)	3	6	9	2	2	4
合計	14	40	54	8	22	30

傾向 (A)：対象言語翻訳文中に原言語の表現が残っている．
傾向 (B)：語句の翻訳や文法に間違いがあり，対象言語翻訳文の翻訳精度が低い．
傾向 (C)：対象言語翻訳文は，文として成立しているが，原文と意味が異なる．

表 7 傾向 (A) による検出失敗例
Table 7 Example sentence of detection failure by cause (A).

原文	それじゃーよろしくおねがいします。
対象言語翻訳文（英語）	それじゃーよろしくおねがいします.
折り返し翻訳文	それじゃーよろしくおねがいします.

この例では，日本語から英語への翻訳に失敗しており，英語であるべき対象言語翻訳文に日本語が含まれている．

表 8 傾向 (B) による検出失敗例
Table 8 Example sentence of detection failure by cause (B).

原文	まあシソが割りと好きです。
対象言語翻訳文（英語）	Oh, I like perillas comparatively.
折り返し翻訳文	ああ、私は比較的にシソが好きである。

この例では，日本語から英語への翻訳において，「まあ」および「割りと」の対訳として文脈上適切でない対訳が選択されている．

表 9 傾向 (C) による検出失敗例
Table 9 Example sentence of detection failure by cause (C).

原文	彼は私の顔をつぶした。
対象言語翻訳文（中国語）	他弄碎了我的脸。
折り返し翻訳文	彼は私の顔をつぶした。

この例では，機械翻訳システムは，日本語の原文を直訳している．しかし，この日本語文には，慣用表現が含まれる．そのため，対象言語翻訳文は文として成立しているものの，原文とは意味の異なる文になっている．

価テキスト B が 6 文），最小値を用いた場合は 30 文中 6 文（評価テキスト A が 4 文，評価テキスト B が 2 文）であった．

5.2.2 傾向 (B)

傾向 (B) には，語句の翻訳がおかしい（単語を直訳している，多義語の選択が間違っている），文法が間違っているなどの原因が含まれる．傾向 (B) による検出失敗例を**表 8** に示す．表 8 に示した英語翻訳文は，日本語原文における程度表現である「まあ」が感動詞「Oh」となっており，また「comparatively」はこのような文において用いないため，翻訳精度が低いと評価された．

傾向 (B) による検出失敗数は，中央値を用いた場合は 54 文中 34 文（評価テキスト A が 6 文，評価テキスト B が 28 文），最小値を用いた場合は 30 文中 20 文（評価テキスト A が 2 文，評価テキスト B が 18 文）であった．

5.2.3 傾向 (C)

傾向 (C) による検出失敗例を**表 9** に示す．表に示した中国語翻訳文は，日本語入力文の「顔をつぶした」をそのまま直訳した文となっている．日本語の文が文字どおり「顔をつぶした」という意味であれば，精度不一致ではない．

しかし，この場合，日本語の文における「顔をつぶす」という表現は，「体面を損なわせる」ということを意味した慣用表現である．そのため，日本語では「体面を損なわせる」ということを意味する文が，中国語では「(物理的に)顔をつぶす」と翻訳されており，入力文の意味と異なると判断され，精度が低いと評価されていた．

中央値を用いた場合は，検出に失敗した精度不一致文 54 文のうち，9 文（評価テキスト A が 3 文，評価テキスト B が 6 文）が該当する．最小値を用いた場合は，検出に失敗した精度不一致文 30 文のうち，4 文（評価テキスト A が 2 文，評価テキスト B が 2 文）が該当する．

6. 考察

本章では，5.2 節で述べた検出に失敗した文の各傾向への対応可能性について議論する．

まず，傾向 (A)（対象言語翻訳文が文として成立しない（原言語の表現が残っている））の対応可能性について検討する．傾向 (A) によって検出に失敗した文については，前処理として，折り返し翻訳の生成前に翻訳失敗しているかどうかを確認することにより，精度の不一致を検出可能である．前処理の適用による検出率の変化を**図 4** に示す．前処理を適用した場合，精度不一致の検出率は，中央値を用いた場合に全体で 48.2%（評価テキスト A は 52.6%，評価テキスト B は 46.9%），最小値を用いた場合に全体で 71.1%（評価テキスト A は 78.9%，評価テキスト B は 68.8%）となった．なお，前処理では原言語の表現の有無を確認しており，原言語の含まれない対象言語翻訳文に前処理を適用しても，誤検出は発生しない．そのため，誤検出率は適用前と同じ値となる．

次に，傾向 (B)（対象言語翻訳文が文として成立しない（語句の翻訳，文法に問題がある））の対応可能性について検討する．傾向 (B) については，翻訳文の文法や語句に問題があるものであり，文としての流暢さや，使われる語句の妥当性を判断することが必要になると考えられる．対象言語翻訳文中の語句の共起確率などを知ることができれば，折り返し翻訳文の精度が高くても，対象言語翻訳文の翻訳精度が低いことを検出できる可能性があると考えられる．

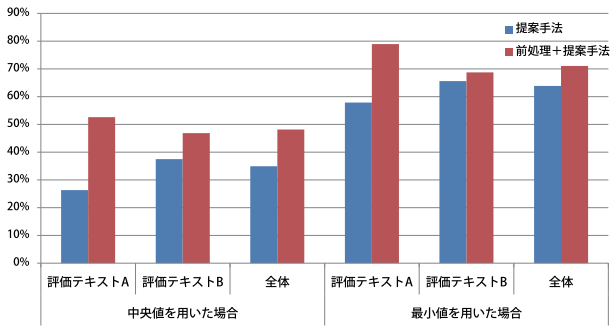


図 4 前処理の適用による検出率の変化
Fig. 4 Detection rate with prior processing.

最後に，傾向 (C)（対象言語翻訳文は，文として成立しているが，原文と意味が異なる）への対応について検討する．傾向 (C) による検出失敗は，対象言語翻訳文自体には問題がないため，検出が容易ではない．対応方法としては，システム側で慣用表現を適切に翻訳できるようにするか，機械翻訳利用者に対し，慣用表現の利用を避けるように促すことで問題の発生を回避するなどの対応が考えられる．

以上のことから，傾向 (A) による検出失敗については，前処理を適用することにより比較的簡単に対応可能であり，中央値を用いた場合は約 48%，最小値を用いた場合は約 71%の精度不一致が検出可能となることを示した．一方，傾向 (B) および (C) については，精度の不一致を単純に検出することは困難であり，今後，これらの傾向を持つ文の精度不一致を検出するための仕組みを検討していく必要がある．

また，代表値として最小値を用いる場合，精度不一致の検出率は向上するが，誤検出率も高くなる．誤検出が発生すると，翻訳精度の高い対象言語翻訳文に関して，問題があるのではないかとユーザに提示することになる．つまり，誤検出は，2 章で述べた，第 2 種の精度不一致（折り返し翻訳文の精度が低いが，対象言語翻訳文の精度は高い）によって発生する問題と同じ問題を引き起こす．2 章で述べたように，第 2 種の精度不一致が発生した場合，ユーザは本来行う必要のない入力文の修正作業を行うことになり，ユーザへの作業負荷が大きくなる可能性がある．一方で，第 1 種の精度不一致のような，意思疎通などの問題の発生にはつながらない．用いる代表値を変えることによる検出率の向上は，誤検出率とトレードオフの関係にある．そのため，正確性の求められる場面では最小値を用い（検出率の向上を優先），短時間での作業が求められる場面では中央値を用いる（誤検出率の低下を優先）など，ユーザの利用目的に応じて，用いる代表値を変更するなどの対応が必要になると考えられる．また，今後，用いる代表値に依存しない検出率の向上が可能かどうか，検討していく必要がある．

今回は，精度不一致の検出が可能かどうかを正確に検証するために，人手による評価結果を用いた．しかし，実際に提案手法を精度不一致検出サービスとして運用する場合は，精度評価を機械的に行う必要がある．これまでに様々な翻訳精度の自動評価手法が提案されている [9], [13], [14], [15], [16]．自動評価手法による判定結果については，人手での評価結果との相関が得られていることが報告されており，今後，精度不一致検出サービスとしての運用も可能であると考えられる．

7. おわりに

機械翻訳を介した多言語間コミュニケーションにおいて，

折り返し翻訳は母語による精度確認手法としての重要な役割を持つ．しかし，折り返し翻訳文の生成においては，2回の翻訳を介するため，対象言語翻訳文と折り返し翻訳文の翻訳精度に不一致が発生する場合がある．本研究では，折り返し翻訳における第1種の精度不一致（対象言語翻訳文が不正確であるが，折り返し翻訳文が正確であるという状況）を検出するために，複数翻訳機を用いた折り返し翻訳の精度不一致検出手法を提案した．提案手法の効果を検証するために，提案手法を用いた折り返し翻訳の精度検証実験を行った．

本研究の貢献は，以下の2点にまとめられる．

(1) 複数翻訳機によって生成された折り返し翻訳文の翻訳精度の代表値として最小値を用いた場合，本提案手法は約64%の精度不一致を検出できることを示した．

(2) 対象言語翻訳文中に原言語表現が残っているかどうかを検証する処理を加えることにより，検出率を約71%へと改善できることを示した．

今後は，本提案手法で検出が困難な精度不一致を検出するための仕組みについて検討を行う．

謝辞 本研究の一部は，日本学術振興会科学研究費基盤研究（B）（22300044）および研究活動スタート支援（23800014）の補助を受けた．

参考文献

[1] Aiken, M.: Multilingual Communication in Electronic Meetings, *ACM SIGGROUP*, Bulletin, Vol.23, No.1, pp.18–19 (2002).

[2] Tung, L.L. and Quaddus, M.A.: Cultural differences explaining the differences in results in GSS: implications for the next decade, *Decision Support Systems*, Vol.33, No.2, pp.177–199 (2002).

[3] 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).

[4] Inaba, R.: Usability of Multilingual Communication Tools, Proceedings, LNCS 4560, pp.91–97 (2007).

[5] Yamashita, N. et al.: Automatic prediction of misconceptions in multilingual computer-mediated communication, *Proc. 11th International Conference on Intelligent User Interfaces*, pp.62–69 (2006).

[6] Yoshino, T., Fujii, K. and Shigenobu, T.: Availability of Web Information for Intercultural Communication, *10th Pacific Rim International Conference on Artificial Intelligence* (*PRICAI 2008*), pp.923–932 (2008).

[7] Morita, D. and Ishida, T.: Designing Protocols for Collaborative Translation, *International Conference on Principles of Practice in Multi-Agent Systems* (*PRIMA-09*), pp.17–32 (2009).

[8] 宮部真衣，吉野　孝：機械翻訳を介したコミュニケーションのための折り返し翻訳の妥当性の検証，電子情報通信学会技術報告，人工知能と知識処理，AI2009-41, pp.65–70 (2010).

[9] Uchimoto, K., et al.: Automatic Rating of Machine Translatability, *10th Machine Translation Summit* (*MT Summit X*), pp.235–242 (2005).

[10] Rapp, R.: The Back-translation Score: Automatic MT Evaluation at the Sentence Level without Reference Translations, *Proc. ACL-IJCNLP 2009 Conference Short Papers*, pp.133–136 (2009).

[11] Walker, K. et al.: *Multiple-Translation Arabic* (*MTA*) *Part 1*, Linguistic Data Consortium, Philadelphia (2003).

[12] Ishida, T.: Language Grid: An Infrastructure for Intercultural Collaboration, *SAINT-06*, pp.96–100 (2006).

[13] Papineni, K., Roukos, S., Ward, T. and Zhu, W.: BLEU: a Method for Automatic Evaluation of Machine Translation, *Proc. 40th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.311–318 (2002).

[14] Denoual, E. and Lepage, Y.：文字単位 BLEU による翻訳自動評価，言語処理学会第11回年次大会発表論文集，pp.522–525 (2005).

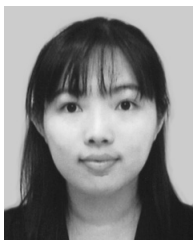
[15] 金山　博，荻野紫穂：翻訳精度評価手法 BLEU の日英翻訳への適用，情報処理学会研究報告，2002-NL-154, pp.131–136 (2003).

[16] 秋葉泰弘, 今村賢治, 隅田英一郎, 中岩浩巳, 山本誠一, 奥乃博：複数の編集距離を用いた口語翻訳文の自動評価，人工知能学会論文誌，Vol.20, No.3, pp.139–148 (2006).

推薦文

折り返し翻訳が正確でありながら翻訳が間違っているという問題を検出するという重要な課題に対して有効なアプローチで取り組み，成果をあげており，推薦論文に値する．

（グループウェアとネットワークサービス研究会
主査　小林　稔）



宮部 真衣 （正会員）

1984年生．2006年和歌山大学システム工学部デザイン情報学科中退．2008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．2011年同大学院システム工学研究科システム工学専攻博士後期課程修了．博士（工学）．現在，東京大学知の構造化センター特任研究員．多言語間コミュニケーション支援，マイクロブログ上の流言拡散防止に関する研究に従事．

吉野 孝 （正会員）

1969年生．1992年鹿児島大学工学部電子工学科卒業．1994年同大学大学院工学研究科電気工学専攻修士課程修了．現在，和歌山大学システム工学部デザイン情報学科准教授．博士（情報科学）．コミュニケーション支援の研究に従事．