RJ-008

折り返し翻訳は本当に役に立たないのか？ ～人間の観点からみた折り返し翻訳の妥当性評価～

Is Back Translation Really Unuseful? Validation of Back Translation from the Perspective of a Checking Method for Users

宮部真衣 † Mai Miyabe
吉野　孝 ‡ Takashi Yoshino

1. まえがき

近年，インターネット上の使用言語の多様化により，ネットワーク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，母語以外の言語により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]．そのため，母語でのコミュニケーションを支援するために，機械翻訳技術を用いた支援が行われている [3]．

近年，機械翻訳技術は急速に進展しているが，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 [4]．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．しかし，原文に対する多言語の翻訳結果を見て，正しく翻訳されているかどうかを判断することは容易ではない．母語のみを用いた多言語の翻訳精度の把握は，折り返し翻訳を利用することにより実現可能である．折り返し翻訳とは，対象言語への翻訳結果を再度母語へと翻訳することである．折り返し翻訳の流れを図 1 に示す．原言語へと再翻訳された折り返し翻訳文は，「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．2 回目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．これまでは，経験的に問題ないと判断され，多言語間コミュニケーションにおける翻訳精度確認手法として折り返し翻訳が利用されていた [3, 5, 6, 7]．しかし，対象言語翻訳と折り返し翻訳の精度不一致が頻繁に発生する場合，折り返し翻訳を精度確認手法として用いることは適切でない．

これまでに，ある翻訳システムが持つ翻訳精度の推定や，複数の機械翻訳システムの翻訳精度の比較を行うという，翻訳システムの性能評価の観点から，翻訳自動評価手法を用いて，折り返し翻訳の利用可能性についての議論が行われている [8, 9]．これらの研究では「折り返し翻訳を精度確認手法として用いることは適切でない」と結論づけられている．これらの研究では，「翻訳システムの性能評価」という観点から，相関があるかどうかのみの確認により，上記の結論を導いている．しかし，折り返し翻訳を精度確認手法として用いる場面としては，上述した翻訳システムの性能評価以外に，コミュニケーションや文書の作成などで，利用する文の翻訳精度を人間が確認する（人間による精度確認）という場面も考えられる．人間が翻訳精度確認のために用いるという観点では，単純に相関の強さによって議論するのは適切でないと考えられる．そのため，人間による精度確認手法としての折り返し翻訳の妥当性については，翻訳システムの性能評価とは別の基準により議論すべきである．

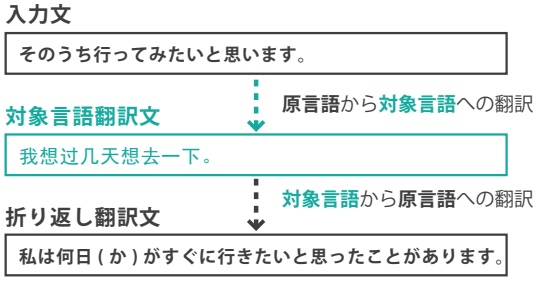

図 1: 折り返し翻訳の流れ

では，人間による精度確認手法としての利用可能性は，どのように議論すべきか．人間が翻訳精度確認手法として利用する場合に最も重要なのは，ある対象言語翻訳結果とその折り返し翻訳結果の翻訳精度に著しい乖離がないということである．そこで，本稿では，翻訳精度の乖離の有無に着目し，人間による精度確認手法としての折り返し翻訳の妥当性について検証する．本研究の特徴は，まず人間の間でも発生しやすい評価結果の乖離を明らかにし，さらに，その乖離に基づいて折り返し翻訳の利用可能性を議論している点にある．

以下，2 章において関連研究について述べる．3 章では人間の観点からみた折り返し翻訳の妥当性評価について述べる．4 章では翻訳精度の主観評価について述べる．5 章で評価結果を示し，6 章で評価結果に関する考察を行う．最後に 7 章で本稿の結論についてまとめる．

2. 関連研究

多言語間コミュニケーションにおいては，母語以外の言語を見て精度を判断することは容易ではない．そのため，著しく精度が異ならない場合は，ユーザが精度確認する際，役に立つ可能性があると考えられる．

これまでに，折り返し翻訳と対象言語翻訳の精度に関する検証が行われている．Somers は，「折り返し翻訳が信頼できるものではない」という機械翻訳の専門家の見解を証明するために，検証実験を行っている [8]．実験の結果，折り返し翻訳はテキストの精度を示すことはできないと述べている．しかし，この研究においては，精度評価において BLEU[10] などの翻訳自動評価手法が用いられている．そのため，翻訳自動評価手法の精度が評価

†東京大学知の構造化センター
‡和歌山大学システム工学部

結果に影響している可能性がある．

Rapp は，BLEU[10] などの翻訳自動評価手法における問題点として，人間の作成した参照訳が必要となる点を挙げ，その問題点の解決のために折り返し翻訳の導入を検討している [9]．また，従来の折り返し翻訳に関する見解の問題点として，翻訳自動評価手法を用いて検証が行われていることを挙げ，BLEU を改良した手法である OrthoBLEU を用いた評価を行っている．Rapp は実験の結果，OrthoBLEU は折り返し翻訳文の評価を改善可能であると述べている [9]．この研究は，従来の見解の問題点として，翻訳自動評価手法を用いた検証を行っているという点を挙げ，人手評価を用いている．しかし，この研究では，OrthoBLEU などの自動評価手法を用いて算出された精度と人手評価との相関を検証し，OrthoBLEU の効果について議論している．つまり，対象言語翻訳および折り返し翻訳のそれぞれに関して，自動評価手法と人手評価との相関を検証したものであり，折り返し翻訳の人手評価結果と対象言語翻訳の人手評価結果との相関については検証を行っていない．

3. 人間の観点からみた折り返し翻訳の妥当性評価

2 章において，これまでに行われてきた折り返し翻訳の妥当性評価に関する研究について述べた．これらの研究では，相関があるかどうかのみの確認により，「折り返し翻訳を精度確認手法として用いることは適切でない」という結論を導いている．「翻訳システムの性能評価」という観点で折り返し翻訳の利用可能性を考えた場合，折り返し翻訳と対象言語翻訳が高い相関をもつ必要がある．

では，人間が翻訳精度確認のために用いる場合も，従来と同様に相関の強さによって議論するのが妥当であろうか？我々は，人間が翻訳精度確認のために用いるという観点では，相関の強さによって議論するのは適切でないと考えた．たとえば，ある翻訳文を見てその精度を判断する場合，どのような判断を下すかは人によって異なる．つまり，ある翻訳文に対する人間の精度評価結果は一意的に定まるものではなく，何らかの範囲として定義されるものであると考えられる．人間による精度確認手法としての利用可能性を議論する場合，このような人間の特性を考慮すべきである．しかし，単に相関の強さを見ただけでは，上記のような特性（各精度評価結果の取りうる値の範囲）は考慮されない．したがって，これまでに行われてきた「翻訳システムの性能評価」という観点での「折り返し翻訳を精度確認手法として用いることは適切でない」という結論を，人間による精度確認手法としての折り返し翻訳の利用可能性に適用するのは早計であり，人間による精度確認手法としての折り返し翻訳の妥当性については，翻訳システムの性能評価とは別の基準により議論すべきである．

では，人間による精度確認手法としての利用可能性は，どのように議論すべきか．人間が翻訳精度確認手法として利用する場合に最も重要なのは，ある対象言語翻訳結果とその折り返し翻訳結果の翻訳精度に著しい乖離がないということである．人間は，自分の翻訳したい文が正しく翻訳されているかどうかを母語で確認するために，折り返し翻訳を用いる．そのため，その対象言語翻訳結果とその折り返し翻訳結果の翻訳精度に著しい乖離が発生していなければ，正しく精度確認が可能である．上述の先行研究では，相関係数をもとに議論が行われているが，テキストセット全体で相関があったとしても，各翻訳結果のペア（対象言語翻訳文および折り返し翻訳文）において翻訳精度の乖離があった場合，精度確認手法としては適切ではない．そこで，本稿では翻訳結果と折り返し翻訳結果の翻訳精度の乖離がどの程度発生しうるかという点から，利用可能性を議論する．

翻訳精度の乖離の発生率を議論するためには，何をもって乖離とするかを定義しなければならない．単純に翻訳精度が「よい」「悪い」という 2 値に分類し，一致しなければ乖離が発生したと判断する手法なども考えられるが，ある翻訳文の精度を複数の人間が評価する場合，すべての人の間で評価結果が一致するとは限らない．そこで，本稿では「人の間でも発生しやすい差異」に着目する．複数の評価者による同一文の評価を行うことにより，人の間で発生しやすい差異と，ほとんど発生しない差異を抽出する．その結果に基づき，単純に 2 値に分類せず，発生しやすい差異を考慮して差異を検証する．また，用途によって，許容される乖離の範囲が変わると考えられるため，差異の閾値を変えた場合の結果を示し，利用者が自分の用途における折り返し翻訳の利用可能性を判断できるようにする．

これまで，様々な研究において人手での翻訳精度評価が行われているものの，代表値を用いて議論を進めており，同じ文に対する評価結果にどの程度の乖離が発生しうるのかは議論されていない．そこで本研究では，まず人間の間で発生しうる評価結果の乖離を明らかにした上で，その乖離に基づいて折り返し翻訳の利用可能性を議論する．

4. 翻訳精度の主観評価

折り返し翻訳の精度確認手法としての妥当性を検証するために，折り返し翻訳文および対象言語翻訳文の翻訳精度について主観評価を行った．

本章では，主観評価実験について述べる．

4.1 評価テキスト

本実験では，評価テキストとして「機械翻訳試験文[1)]」および「チャットにおける発言」の 2 種類の文を用いた．チャットにおける発言は，「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を用いた．評価テキストの一部を表 1 に示す．5 文字以上 44 文字以下の文を各評価テキストからランダムに 200 文[2)]選択し，利用した．以降，機械翻訳試験文をテキストセット 1，チャットにおける発言をテキストセット 2 と呼ぶこととする．

また，原言語の違いによる影響を検証するために，実験用に抽出した日本語の機械翻訳試験文 200 文の英語対訳，中国語対訳，韓国語対訳[3)]を用いて，原言語が英語，

1) NTT Natural Language Research Group, http://www.kecl.ntt.co.jp/icl/mtg/resources/index.php

2) 「5 文字以上 14 文字以下」「15 文字以上 24 文字以下」「25 文字以上 34 文字以下」「35 文字以上 44 文字以下」の文をそれぞれ 50 文選択した．

3) 英語対訳は，機械翻訳試験文内に用意されていたものを用いた．中国語対訳および韓国語対訳については，それぞれ中国語翻訳者，韓国語翻訳者に作成してもらった対訳を用いた．

表 1: 評価テキストの例

テキストセット 1	(1) 私は窓の外を見た。 (2) この小説は想像していたより面白かった。 (3) 梅雨には天気が変わり易いことに留意することが必要だ。 (4) 唯一の違いは彼がコーヒーを飲んだのに対して、彼女が紅茶を飲んだことだ。
テキストセット 2	(5) でもかっこいいですよね。 (6) 私も小さいころはちょっと怖かったです。 (7) ちょっと興味あるんですが屋台でも家でもやったこと無いですねー。 (8) 好きな人はとことん好きな店ですけど、無理な人は絶対嫌って言いますねー。

表 2: 評価ペア数

	評価の組み合わせ	テキストペア数	評価者数
P1	入力文（日本語）とその折り返し翻訳文（日本語）	3600	3
P2	入力文（日本語）とその対象言語翻訳文（英語）	1200	4
P3	入力文（日本語）とその対象言語翻訳文（中国語）	1200	4
P4	入力文（日本語）とその対象言語翻訳文（韓国語）	1200	4
P5	入力文（英語）とその折り返し翻訳文（英語）	600	4
P6	入力文（英語）とその対象言語翻訳文（日本語）	600	4
P7	入力文（中国語）とその折り返し翻訳文（中国語）	600	4
P8	入力文（中国語）とその対象言語翻訳文（日本語）	600	4
P9	入力文（韓国語）とその折り返し翻訳文（韓国語）	600	4
P10	入力文（韓国語）とその対象言語翻訳文（日本語）	600	4

中国語，韓国語の場合の評価を行うこととした．

4.2 使用言語および翻訳システム

本実験では，折り返し翻訳の際の原言語と対象言語の組み合わせを以下の 6 種類とし，精度評価を行う．

［**ペア 1**］原言語：日本語，対象言語：英語
［**ペア 2**］原言語：日本語，対象言語：中国語
［**ペア 3**］原言語：日本語，対象言語：韓国語
［**ペア 4**］原言語：英語，対象言語：日本語
［**ペア 5**］原言語：中国語，対象言語：日本語
［**ペア 6**］原言語：韓国語，対象言語：日本語

本実験では，言語グリッド [11] を介して 3 種類の翻訳システム[4)5)6)]を利用した．なお，折り返し翻訳文の生成については，対象言語翻訳文を生成した場合と同じシステムを用いて行うこととした．

4.3 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）[12] により行った[7)]．適合性評価では，以下の評価基準を用いて，翻訳文が入力文と同じ意味になっているかどうかを比較する．

5： All（同じ意味）

4： Most（文法などに多少問題があるが，大体同じ意味）

3： Much（意味は何となく掴める）

2： Little（雰囲気は残っているが，もとの意味はわからない）

1： None（全く違う意味）

評価者は，日本人大学生 3 名および英語翻訳者 4 名，中国語翻訳者 4 名，韓国語翻訳者 4 名である．表 2 に，評価の組み合わせとテキストペア数を示す．なお，各評価者は全てのテキストの評価を行った．

5. 評価結果

本稿では，各翻訳文の評価にあたり，3 名から 4 名の評価者により評価を行っている．そこで，各翻訳文の精度評価値として，複数評価者による評価結果の中央値を用いて議論を進める．

なお，評価テキストや翻訳システム，使用言語の種類による大きな差異は見られなかったため，本稿では全条件における折り返し翻訳文・対象言語翻訳文のペア（5400 ペア）をまとめて議論する．

5.1 人による評価結果の違い

まず，人による評価結果の違いについて確認を行った．表 2 に示したテキストペアに対し，各評価者数（3 名または 4 名）で評価を行った．図 2 のように，同一文に対する評価結果のペアを抽出し，同一文の評価において共起する評価値の割合を調査した．調査対象となる評価結果のペアは，全部で 54000 ペア[8)]である．評価の結果，

[4)] http://www.kodensha.jp/
[5)] http://translate.google.co.jp/
[6)] http://www.crosslanguage.co.jp/
[7)] Walker らの適合性評価は，2 名以上で行うものである．

[8)] 表 2 の P1 のみ評価者数が 3 名のため，各文に対するペアが 3 通りとなり，P2〜P10 については評価者数が 4 名のため，各文に対するペアが 6 通りとなる．

(A)原文（日本語）	(B)対象言語翻訳文	評価者による(B)の評価結果				
		評価者A	評価者B	評価者C	評価者D	中央値
私は窓の外を見た。	I saw outside the window.	5	5	4	3	4.5

評価値のペア

5（評価者A）－ 5（評価者B）　5（評価者A）－ 4（評価者C）　5（評価者A）－ 3（評価者D）

5（評価者B）－ 4（評価者C）　5（評価者B）－ 3（評価者D）　4（評価者C）－ 3（評価者D）

図 2: 評価結果ペアの抽出例

一部に評価結果の欠損があったため，今回は 53985 ペアのデータで検証を行う．

同一文の評価において共起する評価値の割合を図 3 を示す．図 3 では，評価結果のペアの一方の評価値に対する，もう一方の評価値の発生率を示している．例として，図 3 における「ある評価者がある文に対してつけた評価値」が 5 の場合を説明する．まず，調査対象の 53985 ペアのうち，いずれか一方が 5 のペアを全て抽出した．次に，それらのペアにおける，もう一方の評価値の数を各評価値（1〜5）ごとに集計し，抽出したペア全体に占める割合を調査した．図 3 から，評価値 5 と各評価値の共起率は，評価値 1 が 2%，評価値 2 が 5%，評価値 3 が 14%，評価値 4 が 34%，評価値 5 が 44%であることがわかる．

図 3 より，評価値によって各評価値との共起率は異なるものの，同一文に対する評価値が，各評価者間で必ずしも一致していないことがわかる．例えば，評価結果のペアにおいて評価値が一致する割合を各評価値に関してみてみると，評価値 1 では 39%，評価値 2 が 21%，評価値 3 が 17%，評価値 4 が 17%，評価値 5 が 44%である．

5.2　折り返し翻訳文・対象言語翻訳文の精度評価結果

表 3 に，5400 ペア[9]の折り返し翻訳文・対象言語翻訳文の評価結果を示す．精度不一致の発生率については，次章において議論する．

6.　考察

本章では，5 章で示した結果をもとに，折り返し翻訳の利用可能性について議論する．まず，本稿における精度不一致の定義を述べた後，人による評価結果の違いに基づき，精度不一致判定の許容範囲について定義する．次に，定義した許容範囲に基づき，折り返し翻訳と対象言語翻訳の精度不一致の発生率について考察する．

6.1　本稿における精度不一致の定義

対象言語翻訳文と折り返し翻訳文の精度不一致は，以下の 2 種類が考えられる．

［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**

[9]表 2 で評価する折り返し翻訳文と対象言語翻訳文が対応するため，3600 ペア（P1 と P2,P3,P4），600 ペア（P5 と P6），600 ペア（P7,P8），600 ペア（P9,P10）を合わせて 5400 ペアとなる．

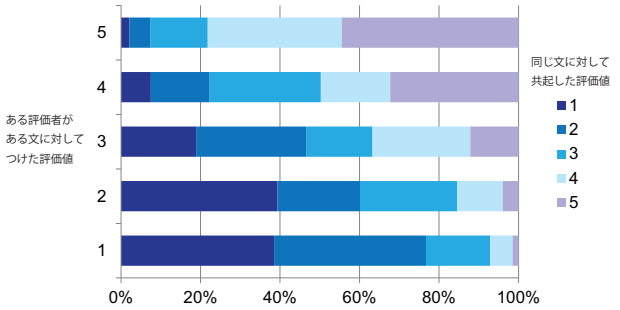

図 3: 同一文の評価において共起する評価値の割合

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ず，意思疎通が困難になる．この状況が多数発生する場合，精度確認の手法として折り返し翻訳を使うことは適切ではない．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可能性のある文を，伝わらない可能性があると判断される．この場合，ユーザは本来不要な修正作業等を行う可能性があるが，第 1 種の精度不一致のような，意思疎通等の問題の発生にはつながらないと考えられる．そこで本稿では，精度確認手法としての妥当性を判断する要素として，第 1 種の精度不一致の発生率を用いる．

6.2　精度不一致の許容範囲

本節では，図 3 に示した同一文の評価において共起する評価値の割合に基づき，精度不一致の許容範囲を定義する．図 3 より，同じ文に対する評価を行っても，人によって評価結果が異なる場合も多いことがわかる．例えば，一方が「5」と評価した場合に，もう一方の評価者が「5」と評価するのが約 45%,「4」と評価するのが約 35%など，必ずしも評価結果は一致しない．

そこで，これらの共起しやすい評価値は，「人の間でも発生しやすい差異」であると考え，各評価値間の共起率に基づいて，精度不一致判定の許容範囲を定義した．表 4 に精度不一致判定の許容範囲の定義を示す．許容範囲は人の評価における評価値の共起率に基づき，以下の 8 段階とする[10]．

[10]評価値が完全に一致した場合を除くと，評価値の共起率の最大値は 39%，最小値は 2%であった．そのため，共起率を 1〜40%の範囲で設定した．

表 3: 対象言語翻訳文と折り返し翻訳文の精度評価結果

		折り返し翻訳の評価結果（中央値）					
		1	2	3	4	5	計
対象言語翻訳の評価結果（中央値）	1	765	531	100	29	21	1446
	2	282	541	221	68	35	1147
	3	152	369	290	155	98	1064
	4	64	212	239	253	185	953
	5	29	76	114	201	370	790
	計	1292	1729	964	706	709	5400

表 4: 人による違いの発生率に基づく精度不一致判定の許容範囲の定義

		折り返し翻訳の評価結果				
		1	2	3	4	5
対象言語翻訳の評価結果	1	MATCH	LEVEL1	LEVEL5	LEVEL7	LEVEL8
	2	第 2 種の精度不一致	MATCH	LEVEL3	LEVEL6	LEVEL8
	3			MATCH	LEVEL3	LEVEL6
	4				MATCH	LEVEL2
	5					MATCH

表中の「MATCH」は，評価値が完全に一致した場合を意味する．各許容範囲は，人の評価における評価値の共起率に基づいて 8 段階で設定した．許容範囲と共起率の対応は以下のとおりである．
LEVEL1：共起率 36〜40%，LEVEL2：共起率 31〜35%，LEVEL3：共起率 26〜30%，LEVEL4：共起率 21〜25%，LEVEL5：共起率 16〜20%，LEVEL6：共起率 11〜15%，LEVEL7：共起率 6〜10%，LEVEL8：共起率 1〜5%

LEVEL1 共起率 36〜40%
LEVEL2 共起率 31〜35%
LEVEL3 共起率 26〜30%
LEVEL4 共起率 21〜25%
LEVEL5 共起率 16〜20%
LEVEL6 共起率 11〜15%
LEVEL7 共起率 6〜10%
LEVEL8 共起率 1〜5%

6.3 精度不一致の発生率

折り返し翻訳と対象言語翻訳の精度が完全に一致した場合のみを精度一致と見なし，それ以外は精度不一致と見なす条件（MATCH）から，許容範囲を LEVEL1〜8 まで拡張していった場合の精度不一致発生率について調査した．図 4 に，許容範囲の拡張に伴う精度不一致発生率の変化を示す．

精度一致条件が MATCH の場合，精度不一致が 26.7% 発生する．LEVEL1〜LEVEL3 までを精度一致と見なした段階で，精度不一致の発生率が 10%を下回る．また，LEVEL1〜LEVEL5 までを精度一致と見なすと，精度不一致の発生率は 5%を下回る．

6.4 折り返し翻訳の精度確認手法としての妥当性

6.3 節において，LEVEL3 までを精度一致と見なすと，精度不一致の発生率は 10%以下，LEVEL5 までを精度一致と見なすと，精度不一致の発生率は 5%を下回ることを示した．

評価値によって共起率は異なるものの，図 3 に示したように，2 名の評価者が全く同じ評価結果を付ける割合はそれほど高くはない．評価値 1 や評価値 5 については，全く同じ評価値の共起率が 40%程度であるが，評価値 3 や評価値 4 については，全く同じ評価値の共起率が

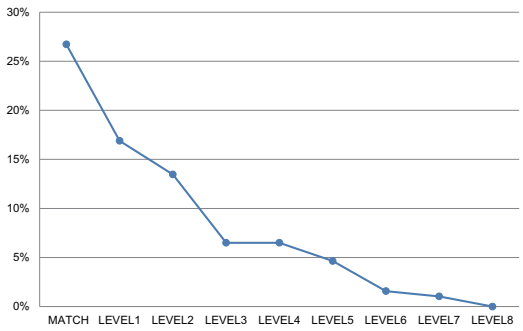

図 4: 許容範囲の拡張に伴う精度不一致発生率の変化

17%程度である．図 3 を見ると，共起率が 15%以下の場合を除いたものを合わせると，各評価値において共起する評価値の 80%以上を含む．そのため，共起率が 15%以下（LEVEL6〜8）の場合は，頻繁に発生するとは言い難いが，共起率 16%以上（LEVEL1〜5）の場合は，同じ言語の全く同じ文を読んだ場合でも発生しうる差異であると考えられる．

したがって，用途によって，許容される不一致の範囲は異なるものの，LEVEL5 までを精度一致と見なした場合の発生率は 5%を下回っており，厳密性が重要視されないような場面では十分利用できる可能性がある．ただし，精度不一致が発生しないわけではないため，利用者に注意を促すなどの対応が必要である．

7. むすび

機械翻訳を介したコミュニケーションにおいて，折り返し翻訳は母語のみを用いた多言語の翻訳精度の把握手法として用いられている．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 回の翻訳を介しており,「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．

先行研究では，翻訳システムの性能評価の観点から，折り返し翻訳の利用可能性についての議論が行われ,「折り返し翻訳を精度確認手法として用いることは適切でない」と結論づけられていた．しかし，これらの研究では，相関があるかどうかのみの確認により，上記の結論を導いており，人間が翻訳精度確認のために用いるという観点からの検証は行われていない．

本稿では，人間による翻訳精度確認の観点から，折り返し翻訳の利用可能性を検証した．用途によって，許容される乖離の範囲が変わると考えられるため,「人の間でも発生しやすい差異」に着目し，対象言語翻訳結果と折り返し翻訳結果の翻訳精度の乖離がどの程度発生しうるかを調査した．本稿の貢献は，以下の点にまとめられる．

1. 同一の文に対する複数評価者の評価結果の共起率を調査し，人間によって発生しうる評価結果の差異（同一文の評価における各評価値の共起率）を明らかにした．

2. 人間による評価結果の差異に基づき，精度不一致の許容範囲を定義し，許容範囲の拡張に伴い精度不一致発生率がどのように変化するかを明らかにした．

本稿で各条件における精度不一致の発生率を示したことにより，用途によって利用者が折り返し翻訳の利用可能性を判断できるようになると考えられる．

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) の助成を受けた．

参考文献

[1] Milam Aiken. Multilingual communication in electronic meetings. *SIGGROUP Bull.*, Vol. 23, pp. 18–19, April 2002.

[2] Lai Lai Tung and M. A. Quaddus. Cultural differences explaining the differences in results in gss: implications for the next decade. *Decis. Support Syst.*, Vol. 33, pp. 177–199, June 2002.

[3] Rieko Inaba. Usability of multilingual communication tools. In *Proceedings of the 2nd international conference on Usability and internationalization*, UI-HCII'07, pp. 91–97, Berlin, Heidelberg, 2007. Springer-Verlag.

[4] Naomi Yamashita and Toru Ishida. Automatic prediction of misconceptions in multilingual computer-mediated communication. In *Proceedings of the 11th international conference on Intelligent user interfaces*, IUI '06, pp. 62–69, New York, NY, USA, 2006. ACM.

[5] Raymond S. Flournoy and Chris Callison-Burch. Secondary benefits of feedback and user interaction in machine translation tools, 2001.

[6] Salvador Climent, Joaquim Moré, Antoni Oliver, Míriam Salvatierra, Imma Sànchez, Mariona Taulé, and Lluïsa Vallmanya. Bilingual newsgroups in catalonia: A challenge for machine translation. *J. Computer-Mediated Communication*, Vol. 9, No. 1, 2003.

[7] Satoshi Sakai, Masaki Gotou, Masahiro Tanaka, Rieko Inaba, Yohei Murakami, Takashi Yoshino, Yoshihiko Hayashi, Yasuhiko Kitamura, Yumiko Mori, Toshiyuki Takasaki, Yoshie Naya, Aguri Shigeno, Shigeo Matsubara, and Toru Ishida. Language grid association: Action research on supporting the multicultural society. In *Proceedings of the International Conference on Informatics Education and Research for Knowledge-Circulating Society (icks 2008)*, ICKS '08, pp. 55–60, Washington, DC, USA, 2008. IEEE Computer Society.

[8] Harold Somers. Round-trip translation: What is it good for? In *Proceedings of the Australasian Language Technology Workshop 2005*, pp. 127–133, Sydney, Australia, December 2005.

[9] Reinhard Rapp. The back-translation score: automatic mt evaluation at the sentence level without reference translations. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pp. 133–136, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[10] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pp. 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[11] Toru Ishida. Language grid: An infrastructure for intercultural collaboration. In *Proceedings of the International Symposium on Applications on Internet*, pp. 96–100, Washington, DC, USA, 2006. IEEE Computer Society.

[12] Kevin Walker, Moussa Bamba, David Miller, Xiaoyi Ma, Chris Cieri, and George Doddington. Multiple-translation arabic (mta) part 1, 2003.