Vol.2009-DD-71 No.4
2009/6/5

折り返し翻訳における中間言語の精度評価

宮部 真衣[†1] 吉野 孝[†2,†3]

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．しかし，中間言語翻訳文と折り返し翻訳文の精度の同等性の検証はこれまでに行われていない．そこで，折り返し翻訳が精度確認のための手法として適切かどうかを検証するために，折り返し翻訳結果と中間言語の翻訳結果の精度評価の比較を行った．評価の結果，以下の知見を得た．(1) 折り返し翻訳文の精度と中間言語翻訳文の精度には正の相関がある．(2)「折り返し翻訳文の精度が高いが，中間言語翻訳文の精度が低い」という精度不一致状況は，1.2%以下であり，今回の評価条件において，意思疎通の問題につながる不一致状況の発生確率は低い．(3) 知見 (1) および (2) より，折り返し翻訳を精度確認手法として利用することによる大きな問題ない．

Accuracy Evaluation of Intermediate Language in Back Translation

Mai Miyabe[†1] and Takashi Yoshino[†2,†3]

In communication using a machine translation, inaccurate translation prevents effective communication between individuals and leads to misunderstandings. Back translation is used to check the accuracy of a sentence translated to a native language. We believe that there is a positive correlation between the accuracy of sentences translated to an intermediate language and that of back-translated sentences. However, this has not yet been verified. We have evaluated the accuracy of back-translated sentences and that of sentences translated to an intermediate language in order to establish the correlation between the two accuracies. We have obtained the following results: (1) There is a positive correlation between the accuracy of sentences translated to an intermediate language and that of back-translated sentences. (2) The occurrence rate of an accuracy mismatch case, wherein a back-translated sentence is accurate but the translated sentence is inaccurate, is less than or equal to 1.2%. (3) Back translation can be used to check the accuracy of a translated sentence.

1. はじめに

近年，世界規模のインターネットの普及に伴った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も高まっている．しかし，一般に多言語を十分に習得することは難しく，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，他言語への翻訳結果を再度母語へと翻訳することである．本稿では，折り返し翻訳を行う際の母語以外の他言語を「中間言語」と呼ぶ．これまでに，中間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は行われていない．

そこで，本稿では，折り返し翻訳が精度確認のための手法として適切かどうかを検証するために，折り返し翻訳結果と中間言語の翻訳結果の精度評価の比較を行う．

以下，2 章において折り返し翻訳利用の課題について述べる．3 章では翻訳精度の主観評価について述べる．4 章で評価結果を示し，5 章で評価結果に関する考察を行う．最後に 6 章でまとめと今後の課題について述べる．

2. 折り返し翻訳利用の課題

折り返し翻訳は，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，機械翻訳を介したコミュニケーションにおいて利用されている[5),6)]．折

†1 和歌山大学大学院システム工学研究科
Graduate School of Systems Engineering, Wakayama University
†2 和歌山大学システム工学部
Faculty of Systems Engineering, Wakayama University
†3 独立行政法人情報通信研究機構 言語グリッドプロジェクト
Language Grid Project, National Institute of Information and Communications Technology

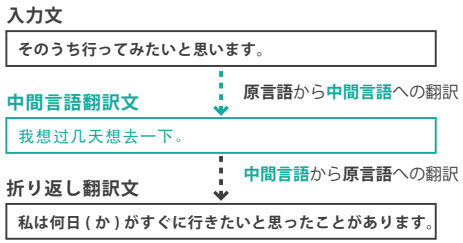

図 1 折り返し翻訳の流れ
Fig. 1 Procedure of back translation.

り返し翻訳の流れを**図 1** に示す．母語へと再翻訳された折り返し翻訳文は,「原言語から中間言語への翻訳」および「中間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,「中間言語から原言語への翻訳」を行うことにより，中間言語の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．折り返し翻訳を精度確認のための手法として用いるには,

(1) 中間言語と折り返し翻訳の精度が正の相関関係にあることが保証されている
(2) 中間言語と折り返し翻訳の精度が大きく異ならない

という条件を満たす必要がある．

これまでは，経験的に中間言語翻訳文と折り返し翻訳文の精度には正の相関があるとし，折り返し翻訳が利用されていたが[7),8)]，中間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は行われていなかった．しかし，折り返し翻訳が精度確認のための手法として適切かどうかを検証する必要がある．そこで本稿では，折り返し翻訳結果と中間言語の翻訳結果の精度評価を行い，母語を用いた翻訳精度の確認手法としての妥当性について議論する．

3. 翻訳精度の主観評価

3.1 検 証 仮 説

本稿では，折り返し翻訳文および中間言語翻訳文の翻訳精度について主観評価を行い，以下の仮説の検証を行う．

[仮説 1]：折り返し翻訳文の精度と中間言語翻訳文の精度には正の相関がある
[仮説 2]：折り返し翻訳文および中間言語翻訳文の精度の相関に関して，入力文の種類の違いによる大きな影響はない
[仮説 3]：折り返し翻訳文および中間言語翻訳文の精度の相関に関して，翻訳システムの違いによる大きな影響はない

表 1 評価に用いたテキストの一部
Table 1 Examples of sentences used in the evaluation.

機械翻訳試験文	(1)	私はデーターベースを検索した。
	(2)	普通は子供が学校へ行く時間だ。
	(3)	大抵の人が帽子をかぶっていた。
	(4)	昔の人は鯨を魚の仲間に入れていた。
	(5)	電球がその回りにほの暗い光を落としていた。
チャットにおける発言	(6)	上を向いてる鳥は 2 つ目っぽいです
	(7)	足をまっすぐ下に伸ばした鳥ですか？
	(8)	背中に何か乗っている犬は 2 匹ですか？
	(9)	前の分を確認するので、ちょっと待ってください。
	(10)	おそらくこれであっていると思います。

(1)〜(5) のテキストは，機械翻訳機能試験文[11)] から 15 文字以上 32 文字以下である文を 100 文選択したものの一部である．
(6)〜(10) のテキストは，機械翻訳を介したチャットによる，図形一致実験[12)] の対話文のうち，15 文字以上 31 文字以下であった文（168 文）の一部である．

仮説 1 が成立し，データ全体を見ると正の相関がある場合においても，文単位では中間言語翻訳文の意味と折り返し翻訳文の意味が同一でない状況（精度不一致状況）が発生する可能性がある．精度不一致状況としては，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，中間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，中間言語翻訳文の精度が**高い**

第 1 種の精度不一致（折り返し翻訳の精度が**高い**が，実際は中間言語の精度が**低い**）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（折り返し翻訳文の精度が**低い**が，中間言語翻訳文の精度が**高い**）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の一つとして，第 1 種の精度不一致の発生率の低さを扱う．翻訳精度評価の結果における精度不一致状況の発生数を測定し，精度確認の手法としての妥当性について議論する．

3.2 翻訳システムと使用言語

本研究では，翻訳システムとして，言語グリッド[9)] を介して高電社の J-Server[10)] を使用した．また，入力言語は日本語，中間言語は中国語および韓国語とし，日本語・中国語翻訳

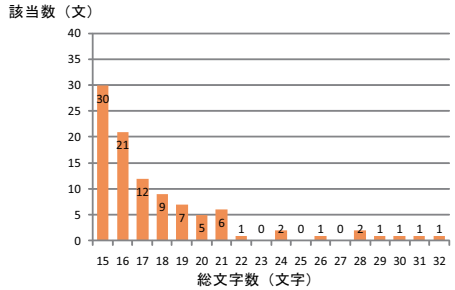

図 2 実験に利用した機械翻訳試験文の総文字数の分布
Fig. 2 Distribution of the total number of characters in the machine translation test set.

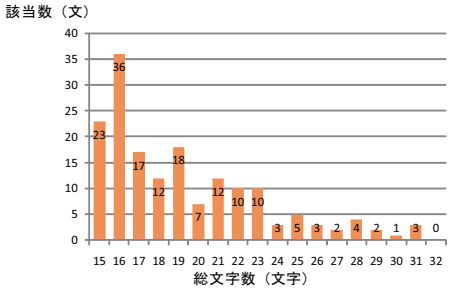

図 3 実験に利用したチャットにおける発言の総文字数の分布
Fig. 3 Distribution of the total number of characters in the sentences used in communication via chat.

および日本語・韓国語翻訳を利用して評価テキストの翻訳を行った．

3.3 評価テキスト

本研究では，入力文の種類による影響を検証するために，評価テキストとして「機械翻訳試験文[11]」および「チャットにおける発言」の 2 種類の文を用いた．評価テキストの一部を**表 1** に示す．また，総文字数の分布を**図 2**，**図 3** にそれぞれ示す．評価テキストは，15 文字以上 32 文字以下の文とし，機械翻訳試験文については，15 文字以上 32 文字以下である文を最初から 100 文選択し利用した．機械翻訳試験文の平均文字数は 17.9 文字，標準偏差は 3.8 文字である．また，チャットにおける発言については，機械翻訳を介したチャットによる図形一致実験[12] の対話文のうち，15 文字以上 31 文字以下（32 文字の文が存在しなかったため）であった 168 文を用いた．チャットにおける発言の平均文字数は 19.2 文字，標準偏差は 4.0 文字である．

3.4 評 価 方 法

折り返し翻訳文および中間言語翻訳文の主観評価は，Walker らの適合性評価（5 段階評価）[13] により行った[⋆1]．以下の 2 組の文について，翻訳文が入力文と同じ意味になっているかどうか比較を行う．

(1) 入力文（日本語）とその折り返し翻訳文（日本語）
(2) 入力文（日本語）とその中間言語翻訳文（中国語または韓国語）

適合性評価の評価基準を以下に示す．

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）

評価は，並べられた 2 文を見て，1 組の文に対して 30 秒以内で評価するものとした．

評価者は，日本人大学生 3 名および日本語の読み書きが可能な中国人留学生 4 名，韓国人留学生 3 名であり，日本人大学生は (1) の文の比較を，中国人留学生および韓国人留学生は (2) の文の比較をそれぞれ行う．

なお，韓国語の評価は，機械翻訳試験文のみを対象として行った．これは，評価テキスト「チャットにおける発言」を抽出した機械翻訳を介したチャットによる図形一致実験[12] では，中間言語として中国語を用いており，韓国語の中間言語翻訳ログが存在しなかったためである．

4. 評 価 結 果

4.1 折り返し翻訳文と中間言語翻訳文の精度の相関

今回用いた評価基準は 5 段階評価である．3.4 節の評価基準より，評価差が 1 の場合は大きな差ではないと見なすこととした．評価差が 2 以上の場合を考えると，例えば，一方が 4（文法などに多少問題があるが，大体同じ意味），もう一方が 2（雰囲気は残っているが，もとの意味はわからない）の場合，精度は大きく異なる．そこで，**評価値の差の絶対値が 2 以上である場合，精度が一致していないと判断することとする**．

⋆1 Walker らの適合性評価は，2 名以上で行うものである．

表 2 各評価テキストにおける平均精度評価値
Table 2 Average evaluated accuracy on each test set.

		機械翻訳試験文（中国語）	チャットにおける発言（中国語）	有意確率 [1]
平均精度（標準偏差）	折り返し翻訳文	2.7(1.2)	3.1(1.2)	0.003*
	中間言語翻訳文	3.6(1.3)	3.5(1.1)	0.259
有意確率 [2]		0.000*	0.000*	
評価値の差の絶対値		1.2	0.9	

*：有意差あり $p<0.01$
1：マンホイットニー検定を利用
2：ウィルコクスンの符号付き順位検定を利用

表 3 各翻訳システムにおける平均精度評価値
Table 3 Average evaluated accuracy on each translation system.

		機械翻訳試験文（中国語）	機械翻訳試験文（韓国語）	有意確率 [1]
平均精度（標準偏差）	折り返し翻訳文	2.7(1.2)	3.8(1.2)	0.000*
	中間言語翻訳文	3.6(1.3)	4.1(0.9)	0.000*
有意確率 [1]		0.000*	0.071	
評価値の差の絶対値		1.2	0.9	

*：有意差あり $p<0.01$
1：ウィルコクスンの符号付き順位検定を利用

表 4 折り返し翻訳文と中間言語翻訳文の精度の相関係数
Table 4 Correlation coefficient between the accuracy of back-translated sentence and that of translated sentence.

	相関係数	有意確率
機械翻訳試験文（中国語）	0.478	0.000
チャットにおける発言（中国語）	0.585	0.000
機械翻訳試験文（韓国語）	0.432	0.000

評価者による精度評価結果を**表 2** および**表 3** に，折り返し翻訳文と中間言語翻訳文の精度の相関係数を**表 4** にそれぞれ示す．表 2，表 3 における「機械翻訳試験文（中国語）」は同一のデータである．

表 2 より，機械翻訳試験文，チャットにおける発言のどちらの評価テキストについても，折り返し翻訳文と中間言語翻訳文の精度評価値の平均には有意差が見られる．表 3 より，韓国語翻訳については有意差は見られなかった．評価値の差の絶対値を確認したところ，機械翻訳試験文は平均 1.2，チャットにおける発言は平均 0.9，韓国語翻訳は平均 0.9 となってお

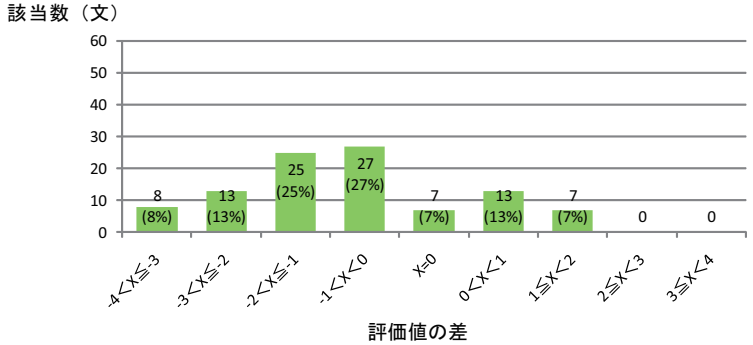

図 4 折り返し翻訳文と中間言語翻訳文の差の分布（機械翻訳試験文，中国語）
Fig. 4 Distribution of difference between the evaluated accuracy of back-translated sentence and that of Chinese-translated sentence in the machine translation test set.

表 5 精度不一致状況の発生率
Table 5 Occurrence rate of accuracy mismatch cases.

	機械翻訳試験文（中国語）	チャットにおける発言（中国語）	機械翻訳試験文（韓国語）
第 1 種の精度不一致（折り返し翻訳文は高精度だが，中間言語翻訳文は低精度）	0 文（0%）	2 文（1.2%）	0 文（0%）
第 2 種の精度不一致（折り返し翻訳文は低精度だが，中間言語翻訳文は高精度）	21 文（21.0%）	13 文（7.7%）	13 文（13.0%）

り，多少精度は異なるものの，全体としては不一致の状態にはなっていないと考えられる．なお，文ごとに見た場合の精度不一致状況の発生数については，次節で述べる．

また，表 4 より，機械翻訳試験文（中国語），チャットにおける発言および機械翻訳試験文（韓国語）における折り返し翻訳文と中間言語翻訳文の精度の相関係数は，それぞれ 0.478，0.585，0.432 となっており正の相関が見られる．したがって，**[仮説 1]：折り返し翻訳文の精度と中間言語翻訳文の精度には正の相関がある**が成立する．

4.2 精度不一致状況

4.1 節で述べたように，本稿では評価値の差の絶対値が 2 以上である場合，精度が一致していないと判断する．3.1 節で述べた 2 種類の精度不一致状況への該当条件は，以下の通り

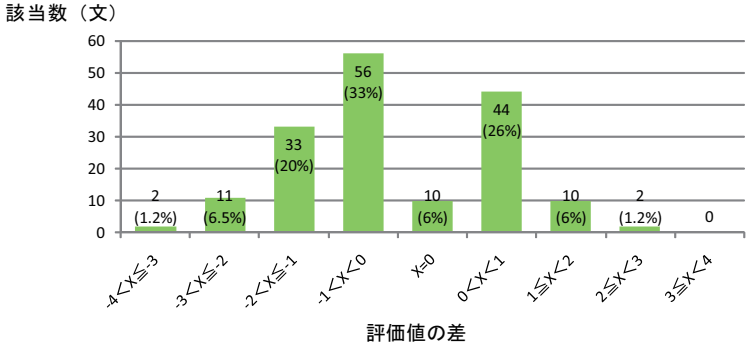

図 5 折り返し翻訳文と中間言語翻訳文の差の分布（チャットにおける発言，中国語）

Fig. 5 Distribution of difference between the evaluated accuracy of back-translated sentence and that of Chinese-translated sentence in the sentences used in communication via chat.

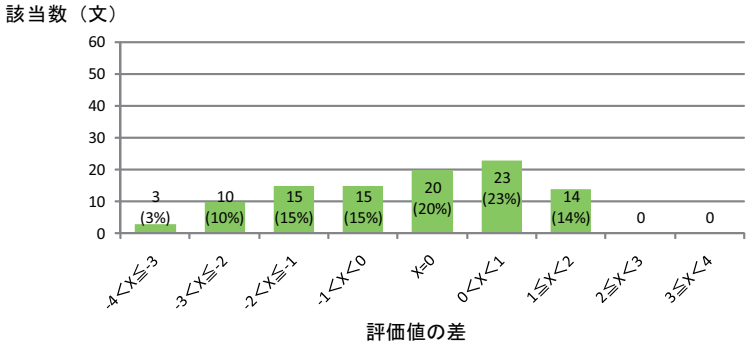

図 6 折り返し翻訳文と中間言語翻訳文の差の分布（機械翻訳試験文，韓国語）

Fig. 6 Distribution of difference between the evaluated accuracy of back-translated sentence and that of Korean-translated sentence in the machine translation test set.

である．

［第 1 種の精度不一致］：折り返し翻訳文の精度評価値－中間言語翻訳文の精度評価値 ≥ 2

［第 2 種の精度不一致］：折り返し翻訳文の精度評価値－中間言語翻訳文の精度評価値 ≤ -2

そこで，上記の状況の発生数の確認を行った．折り返し翻訳文と中間言語翻訳文の評価値差の分布を**図 4**，**図 5**，**図 6** に示す．また，折り返し翻訳文および中間言語翻訳文の精度不一致状況の発生数を**表 5** に示す．

図 4，図 5，図 6 より，評価値の差が 2 以上あるいは-2 以下である文の数は少ないことがわかる．また，表 5 より，第 1 種の精度不一致の発生率は，機械翻訳試験文（中国語）では 0 文（0%），チャットにおける発言では 2 文（1.2%），機械翻訳試験文（韓国語）では 0 文（0%）であった．一方，第 2 種の精度不一致は，機械翻訳試験文で 21 文（21%），チャットにおける発言で 13 文（7.7%），機械翻訳試験文（韓国語）で 13 文（13.0%）発生していた．

5. 考　　察

5.1 入力文の種類の違いによる影響

表 4 に示すように，機械翻訳試験文における精度の相関係数は 0.478，チャットにおける発言における精度の相関係数は 0.585 であった．評価テキストによる違いがあるかどうかを検証するために，相関の差の検定[14] を行った．帰無仮説を「2 つの母相関係数は等しい」とし，有意水準 5%で検定を行ったところ，$P = Pr\{\chi^2 \geq 1.37\} > 0.05$ となり，帰無仮説は棄却されない．

本稿では，**［仮説 2］：折り返し翻訳文および中間言語翻訳文の精度の相関に関して，入力文の種類の違いによる大きな影響はない**という仮説を立てた．今回の評価結果では，帰無仮説「2 つの母相関係数は等しい」は棄却されておらず，2 つの母相関係数は異なるとはいえない．したがって，今回の評価結果では，入力文の種類の違いによる，精度の相関への大きな影響はないと考えられる．

5.2 翻訳システムの違いによる影響

今回の評価では，翻訳システムの違いによる検証を行うために，異なる中間言語（中国語または韓国語）を用いた場合の，機械翻訳試験文に関しての精度評価を行った．表 4 より，中間言語が中国語の場合の精度の相関係数は 0.478，韓国語の場合の精度の相関係数は 0.432 であった．翻訳システムによる違いがあるかどうかを検証するために，帰無仮説を「2 つの母相関係数は等しい」とし，有意水準 5%で相関の差の検定を行ったところ，$P = Pr\{\chi^2 \geq 0.16\} > 0.05$ となり，帰無仮説は棄却されない．

本稿では，**［仮説 3］：折り返し翻訳文および中間言語翻訳文の精度の相関に関して，翻訳システムの違いによる大きな影響はない**という仮説を立てた．今回の評価結果では，帰無仮説「2 つの母相関係数は等しい」は棄却されておらず，2 つの母相関係数は異なるとはいえない．したがって，今回の評価結果では，翻訳システムの種類の違いによる，精度の相関への大きな影響はないと考えられる．

5.3 折り返し翻訳の精度確認手法としての妥当性

2 章において，精度確認手法としての妥当性を示すための条件として，

(1) 中間言語と折り返し翻訳の精度が正の相関関係にあることが保証されている

(2) 中間言語と折り返し翻訳の精度が大きく異ならない

を挙げた．

今回の評価では，**[仮説 1]：折り返し翻訳文の精度と中間言語翻訳文の精度には正の相関がある**が成立した．精度評価値の差については，機械翻訳試験文は平均 1.2，チャットにおける発言は平均 0.9，韓国語翻訳は平均 0.9 であり，多少精度は異なるものの，全体としては不一致の状態にはなっていない．また，妥当性に大きく影響する第 1 種の精度不一致（折り返し翻訳文の精度が**高い**が，中間言語翻訳文の精度が**低い**）の発生率は，チャットにおける発言において 1.2%であったが，機械翻訳試験文（中国語，韓国語）では 0%となっており，第 1 種の精度不一致が発生する可能性は低いと考えられる．

したがって，今回の評価結果では，精度確認手法として折り返し翻訳を利用しても問題ないと考えられる．

6. おわりに

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，「原言語から中間言語への翻訳」および「中間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ため，「中間言語から原言語への翻訳」を行うことにより，中間言語の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．しかし，中間言語翻訳文と折り返し翻訳文の精度の同等性についてはこれまでに検証されていない．

本稿では，折り返し翻訳が精度確認のための手法として適切かどうかを検証するために，15 文字以上 32 文字以下の評価テキストを用いて，折り返し翻訳結果と中間言語の翻訳結果の精度評価の比較を行った．評価の結果，以下の知見を得た．

(1) 折り返し翻訳文の精度と中間言語翻訳文の精度には正の相関がある．

(2) 「折り返し翻訳文の精度が**高い**が，中間言語翻訳文の精度が**低い**」という精度不一致状況の発生率は，1.2%以下であり，今回の評価条件において，意思疎通の問題につながる不一致状況の発生確率は低い．

(3) (1) および (2) より，折り返し翻訳を精度確認手法として利用することによる大きな問題ない．

謝辞 本研究は，日本学術振興会科学研究費基盤研究 (B)(19300036) の補助を受けた．

参 考 文 献

1) Aiken, M.: Multilingual Communication in Electronic Meetings, ACM SIGGROUP, Bulletin, 23, 1, pp.18-19 (2002).
2) Tung, L.L. et al.: Cultural differences explaining the differences in results in GSS: implications for the next decade, Decision Support Systems, 33, 2, pp.177-199 (2002).
3) Inaba, R.: Usability of Multilingual Communication Tools, Proceedings, Lecture Notes in Computer Science 4560, pp.91-97 (2007).
4) Yamashita, N. et al.: Automatic prediction of misconceptions in multilingual computer-mediated communication, Proc. the 11th international conference on Intelligent user interfaces, pp.62-69 (2006).
5) 坂本知子，野村早恵子，石田亨，井佐原均，小倉健太郎，林良彦，石川開，小谷克則，島津美和子，介弘達哉，畠中伸敏，富士秀，船越要：機械翻訳システムに対する利用者適応の分析 －異文化コラボレーションを目指して－，情報処理学会研究報告，2003-ICS-135，pp.125-130（2004）．
6) Miyabe, M. et al.: Effects of Repair Support Agent for Accurate Multilingual Communication, Proceedings, Lecture Notes in Computer Science 5351, pp.1022-1027 (2008).
7) 藤井薫和　他：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）．
8) 森田大翼　他：共同翻訳のためのプロトコルと支援システムの開発，FIT2008 情報科学技術フォーラム，第 3 分冊，pp.417-420（2008）．
9) Ishida, T.: Language Grid: An Infrastructure for Intercultural Collaboration, IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06), pp.96-100(2006).
10) KODENSHA，http://www.kodensha.jp/
11) NTT Natural Language Research Group, http://www.kecl.ntt.co.jp/icl/mtg/resources/index.php
12) 宮部真衣　他：折り返し翻訳を用いた翻訳リペアのチャットコミュニケーションへの影響，情報処理学会研究報告，2009-GN-070，pp.109-114（2009）．
13) Walker, K. et al.: Multiple-Translation Arabic (MTA) Part 1, Linguistic Data Consortium, Philadelphia (2003).
14) 森敏昭，吉田寿夫：心理学のためのデータ解析テクニカルブック，北大路書房（1990）．