

機械翻訳を用いた高精度な文章作成のための 翻訳精度表示の影響

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において、翻訳精度は話者間の相互理解に影響する。翻訳リペアを行うことで、ユーザは翻訳精度を向上することができる。しかし、翻訳リペアは、ユーザが文章の精度が低いと判断した段階で、初めて行われる作業であるため、ユーザの不正確な判定により、翻訳リペアの効果が得られない場合がある。この問題を解決するために、本稿では翻訳リペアにおける翻訳精度表示を提案する。翻訳自動評価手法を用いて精度を測定し、精度を%表示と5段階表示を併用した手法によって表示する。翻訳精度表示を行うことによる、ユーザの不正確判定に対する効果を検証するために、テキスト修正実験を行った。実験の結果、以下の知見を得た。(1) 本実験においては、精度表示を行うことによる不正確判定の減少効果は見られなかった。ただし、自動評価を用いたため、自動評価の失敗により不正確判定につながった場合もあると考えられる。今後自動評価の精度が向上することにより、不正確判定の減少効果につながる可能性があると考えられる。(2) アンケートの結果、精度表示は役に立つ可能性があるものの、自動精度評価の精度に問題があり、改善する必要があることがわかった。

Verification of Effects of Translation Accuracy Indication for Generating Accurate Sentences using Machine Translation

MAI MIYABE^{†1} and TAKASHI YOSHINO^{†2,†3}

Translation accuracy affects mutual understanding between people in communication using machine translation. People can improve the accuracy of translated sentences by undertaking translation repair. Translation repair is typically performed only when a user thinks that his/her message is inaccurate. Inaccurate judgments made by users in this regard cause the debasement of translation accuracy. In order to solve this problem, we propose the method that provides indications of translation accuracy to users. In this method, we measure the accuracy of translated sentences using an automatic evaluation

method, and provide it using the indications of a percentage and a five-point scale. We verified the effects of the indications for reduction of inaccurate judgments. The following conclusions were drawn from experimental results: (1) The indications of translation accuracy did not affect the inaccurate judgments of users significantly. However, in this experiment, automatic-evaluated values were not always accurate. We think that wrong automatic-evaluated values might have led some inaccurate judgments. If the accuracy of an automatic evaluation method is improved, the indications of translation accuracy may reduce the inaccurate judgments. (2) From the result of the questionnaire, the indication of translation accuracy was useful for the judgment of translation accuracy. However, the accuracy of an automatic evaluation method was inaccurate occasionally. We need to use a more accurate method of an automatic evaluation.

1. はじめに

近年、世界規模のインターネットの普及により、電子メールや掲示板、チャットなどのコミュニケーションツールが広く利用されている。また、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)}。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),(4)}。

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。精度の低い文章は話者間の相互理解を困難にし、円滑なコミュニケーションの妨げとなる。したがって、円滑にコミュニケーションを行うためには、精度の高い文章を作成しなければならない。翻訳精度を向上させるために、入力文章を書き換えていくことを「翻訳リペア」と呼ぶ⁽⁵⁾。

折り返し翻訳^{*1}を用いた翻訳リペアによって得られる翻訳結果の精度検証実験から、翻訳

†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

*1 折り返し翻訳とは、他言語への翻訳結果を再度原言語へと翻訳することである。

リペアを行うことによって翻訳精度が改善できることが確認されている⁵⁾。一方、翻訳リペアは、文章の精度が低いとユーザが判断した段階で初めて行われる作業である。そのため、ユーザの不正確な翻訳精度の判断（不正確判定）により、低精度なメッセージの未修正や、十分にリペアがされていない状態での修正終了などが発生している。円滑なテキストコミュニケーションを実現するには、低精度なメッセージが用いられないようにする必要があり、ユーザの「不正確判定」を減少させるための仕組みが必要となる⁶⁾。これまでに、チャットコミュニケーションにおいて、不正確判定による低精度なメッセージ送信を防ぐために、精度判定に基づくメッセージ送信拒否の適用可能性についての検証が行われている⁷⁾。精度判定に基づくメッセージ送信拒否は、コミュニケーションにおける低精度なメッセージの利用を防ぐことはできるが、ユーザの不正確判定自体を減少させることは難しいと考えられる。そのため、ユーザの不正確判定を減少させる仕組みが必要である。

不正確判定はユーザ自身の精度の判断基準が他のユーザと異なることが原因で発生していると考えられる。客観的な精度を表示することで、不正確判定を減少させることができる可能性がある。本研究では、ユーザの不正確判定を減少させるために、翻訳精度をユーザに表示する仕組みを検討する。具体的には、自動翻訳評価を用いて、ユーザに対して翻訳精度表示を行う。本稿では、機械翻訳を用いた高精度な文書作成を支援するための手法として、翻訳精度を表示することによる影響の検証を行う。

2. 機械翻訳利用におけるユーザの不正確な翻訳精度判定

我々は、これまでに行った翻訳リペア実験において、精度が十分でなく、翻訳リペアすべき文章に対して、被験者は修正不要であると判断する（不正確判定）場合があることを明らかにした⁶⁾。我々の実施した実験においては、ユーザの不正確判定率^{*1}は平均7%、最大23%であった。しかし、この不正確判定率は、ユーザが全く修正をしなかった文の数であり、翻訳リペアを行ったが、不十分な状態で修正を終えた文の数が含まれていない。そのため、低精度な文の数はさらに多く、不正確判定率はより高い⁶⁾。

コミュニケーションにおいて機械翻訳を適切に利用するためには、低精度なメッセージの利用を防ぐことが必要であると考えられる。低精度なメッセージの利用を防ぐためには、以下の対応が必要になると考えられる。

(1) 低精度な状態におけるメッセージの作成終了判断（不正確判定）の防止

*1 本来修正すべき文を、被験者が修正不要であると判断した割合を不正確判定率と呼ぶ。

(2) コミュニケーションにおける低精度なメッセージの利用の防止

どちらの対応についても、入力メッセージと折り返し翻訳文が同じ意味を持っているかどうかを判定する必要がある^{*2}。この文意一致判定を、本稿では「精度判定」と呼ぶ。

不正確判定は、ユーザが「低精度である」と判断する基準が低いことによって発生すると考えられる。そのため、客観的な指標として翻訳精度を提示し、翻訳精度に関する気付きを与えることによって、ユーザの判断に影響を与え、不正確判定の減少ができる可能性がある。

本稿では、翻訳自動評価手法を用いて折り返し翻訳結果の精度判定を行う。また、判定結果をユーザに提示することの、ユーザの不正確判定に対する影響を検証する。

3. 折り返し翻訳文の翻訳精度表示

メッセージ作成時に翻訳精度を表示するためには、作成したメッセージの翻訳精度を測定する必要がある。そこで、本研究では、翻訳自動評価手法を用いて精度を測定する。

これまでに、機械翻訳システムの評価のために、BLEU⁸⁾ や NIST⁹⁾ など、様々な翻訳自動評価手法の提案や評価がなされている¹⁰⁾⁻¹²⁾。これらの評価手法は基本的には対象言語への翻訳結果とバイリンガルが作成した対象言語の参照訳とを比較するものであるが、折り返し翻訳結果と原文の比較による手法も提案されている¹³⁾。本研究では、折り返し翻訳を用いたメッセージ作成を想定しており、後者の手法を利用して作成したテキストの精度評価を行うことが可能である。

3.1 表示手法

我々は、効果的な精度の表示手法を調べるための予備実験を行った¹⁴⁾。この実験においては、以下の3種類の表示手法を用いて、文章修正実験を行った。

手法 A %表示 (0% ~ 100%)

手法 B 5段階表示^{*3} (完全に翻訳されている / 流暢ではないが、問題なく意味を読み取れる / 何となく意味はつかめる / 雰囲気は残っているが、もとの意味はわからない / 正しく翻訳されていない)

手法 C 3段階表示 (良い / どちらとも言えない / 悪い)

実験の結果、3つの表示手法では、手法 B (5段階表示) が最も高い評価を、手法 A (%表示)

*2 ユーザは折り返し翻訳を見ながら翻訳リペアを行うため、本論文では精度判定において、対象言語に翻訳された文ではなく、折り返し翻訳文を用いる。

*3 5段階表示における表現については、翻訳精度の人手評価で用いられている Walker らの適合性評価¹⁵⁾をもとに作成した。

表 1 自動評価値および 5 段階表示の対応

Table 1 Correspondence table between an automatic-evaluated value and a five-point scale.

自動評価値	5 段階表示
$1 \leq Value < 2$	正しく翻訳されていない
$2 \leq Value < 3$	雰囲気は残っているが、もとの意味はわからない
$3 \leq Value < 4$	何となく意味はつかめる
$4 \leq Value < 5$	流暢ではないが、問題なく意味を読み取れる
$Value = 5$	完全に翻訳されている

示)が2番目に高い評価を得た。しかし、どちらの手法についてもデメリットが指摘されている。5段階表示は「尺度が明確である」が、「細かい差がわからない」という評価を得ていた。また、%表示は「精度の変化量がわかりやすい」が、「満足する基準を定めにくい」という評価を得た。「%表示」は「5段階表示」のデメリットを補うことができると考えられ、2つの手法を併用することによってより判断しやすい表示となる可能性があることを示した。

そこで、本稿では、手法 A と手法 B を併用した精度表示を行い、その効果の検証を行う。

3.2 精度判定方法

本研究では、翻訳精度の自動評価指標として BLEU (Bilingual Evaluation Understudy)⁸⁾ を用いる。BLEU スコアに基づき、提案する 3 種類の精度表示を行う。

BLEU は、n-gram の一致度を用いる自動評価手法である。n の値が小さい場合は単語を重視した指標となり、n の値が大きくなるほど文の流暢さを重視した指標となる。本研究では、文の流暢さではなく、意味が通じるかどうかを重視し、単語を重視した指標を用いることとする。そこで、1-gram の BLEU スコアを算出し、精度として用いる。以下の計算式^{*1}により BLEU スコアを算出する (なお、c は折り返し翻訳文中の形態素数^{*2}、r は原文中の形態素数である。)

$$BLEU = BP \times \exp(\log P) \quad (1)$$

$$BP = \begin{cases} 1 & (c \geq r) \\ \exp(1 - r/c) & (c < r) \end{cases} \quad (2)$$

$$P = \frac{\text{原文と折り返し翻訳との形態素の共有数}}{c} \quad (3)$$

*1 Papineni らの提案した BLEU スコアの計算式^{8),10)} の n-gram を 1 としたものである。

*2 本研究では、形態素解析器 Mecab¹⁶⁾ を用いて文を形態素に分割する。

BLEU スコアは 0 から 1 の間の値をとる。そこで、手法 A については BLEU スコアを表示することとした。

手法 B, C については、BLEU スコアを 5 段階評価値に対応させることにより、自動的に算出できるようにする。5 段階評価を用いた人手による翻訳精度評価の結果 (人手評価値) と BLEU スコアとの相関の検証結果¹⁷⁾ に基づき、回帰式を求めた。5 段階評価値は以下の計算式により算出する。

$$Value = 2.856 \times BLEU + 1.183 \quad (4)$$

BLEU スコアに基づいて自動で算出される 5 段階評価値と 5 段階表示の対応付けを表 1 に示す。

4. 検証実験

機械翻訳を用いた多言語の文章作成において翻訳精度を表示することによる影響を検証するために、折り返し翻訳を用いたテキスト修正実験を実施した。本実験においては、精度表示を行うことにより、不正確判定が減少するかどうかを検証する。

実験の被験者は、大学生および大学院生 12 名である。本実験ではメッセージ入力が必要であるため、キーボード入力に慣れている学生を被験者とした。被験者の年齢は 18 歳から 23 歳 (平均 21 歳) である。

4.1 実験内容

修正するテキストとして、会話表現データベース¹⁸⁾ のうち、20 文字から 30 文字のテキストを 60 文を用いた^{*3}。

本実験では、不正確判定の減少効果について検証するために、修正が必要である文について分析を行う。そこで、会話表現データベースにおける全トピックのテキストの順序をランダムにした後、該当文字数であるテキストのうち、分析対象となる修正が必要である文を 40 文、分析対象外となる特に修正する必要のない文を 20 文選択した。利用したテキストの

*3 先行研究により、15 文字未満の文は高精度の翻訳結果が得られる可能性が特に高い傾向があることが分かっている¹⁹⁾。また、これまでに行った実験では、20 文字以下の文についても高精度であり、修正が不要な文章が多かったため、本実験では 20 文字を実験対象の下限とする。一方で、機械翻訳システムでは、50 文字以上の日本語文の解析は非常に困難であり、80 文字以上の文については解析がほとんど失敗すると指摘されている^{20),21)}。現在、機械翻訳に必要な機能の改善が進んでいるが、十分な精度は得られていない²²⁾。また、これまでの実験において 32 文字以下の文については、リペア作業の適用可能性を確認できているため、本実験における実験対象の上限を 30 文字とする。以上のことから、本実験では 20 文字以上 30 文字以下の日本語文について翻訳リペア作業を行うこととした。

表 2 実験に用いたテキストの一部
Table 2 Examples of sentences used in the experiment.

(1)	あなたの名前と住所をここに書いてください。
(2)	あの車が急に角から飛び出してきたんです。
(3)	あまりそうじなどが行き届いていませんでした。
(4)	お泊まりになっているホテルでは食事はしないほうがいいですよ。
(5)	ステーションワゴンを一週間借りたいのですが。

これらのテキストは、会話表現データベース¹⁸⁾ から 20 文字以上 30 文字以下である文を 60 文選択したもの的一部である。

一部を表 2 に示す。

4.1.1 実験条件

仮説を検証するために、本実験では以下の 2 種類の実験条件下において実験を行う。

条件 1 精度表示なし

条件 2 精度表示あり

実験では、それぞれの条件下において 30 文ずつ修正作業を行うこととした。

4.1.2 実験ツール

本実験で用いた実験ツールの画面を図 1 に示す。入力エリア (図 1 (3)) に入力した文の折り返し翻訳結果が、折り返し翻訳表示エリア (図 1 (2)) に表示される。実験条件 2 の場合のみ、折り返し翻訳表示エリアに表示された折り返し翻訳結果の後に判定した精度が赤字で表示される (図 1 (1))。

翻訳システムは、言語グリッド²³⁾ を介して高電社の J-Server²⁴⁾ を使用した。また、翻訳の言語ペアについては、入力言語を日本語、対象言語を中国語とした^{*1}。BLEU スコア算出における単語抽出においては、形態素解析器 Mecab¹⁶⁾ を用いた。

4.1.3 実験手順

実験の手順を以下に示す。

- (1) 折り返し翻訳結果を見ながら、原文と同じ意味になるように入力文を修正する。
- (2) 折り返し翻訳結果が同じ意味だと判断したら、次の文の修正を行う。
- (3) (1)~(2) の手順を 30 文行う。
- (4) 実験条件を切り替え、1~3 の手順を行う。

*1 テキストを修正する必要があるのは、テキストの精度が低い場合である。対象言語を英語、中国語、韓国語とし、予備実験を行った結果、中国語の翻訳が最も精度が低かった。そのため、本稿では翻訳対象言語を中国語とし、実験を行った。

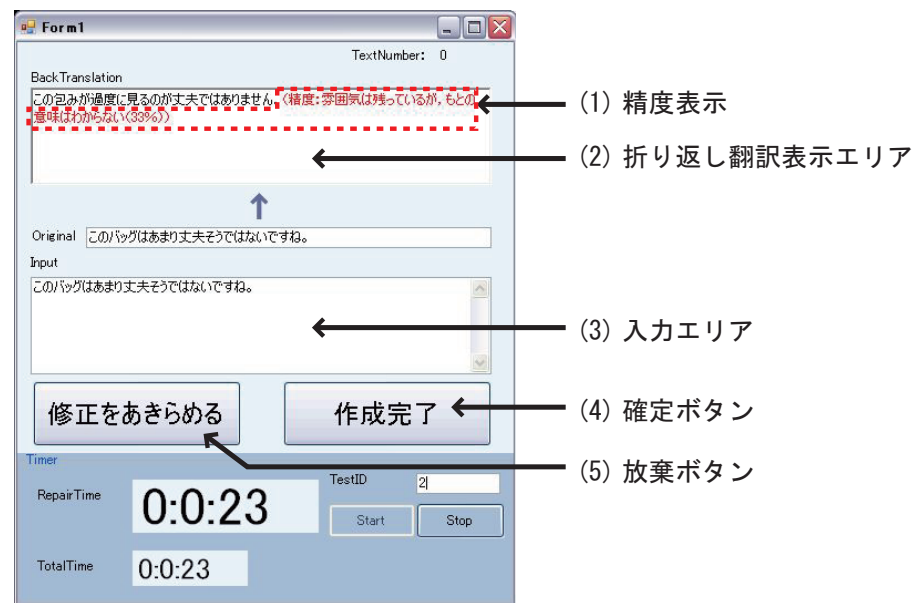


図 1 実験ツールの画面例

Fig. 1 Screenshot of an experimental tool for translation repair.

何も修正していない状態における折り返し翻訳結果が原文と同じ意味になっていると被験者が判断した場合は、リペア作業を行わず、そのまま確定可能とした。また、5 分程度リペア作業を行っても同じ意味の文へと修正ができない場合、修正を放棄させることとした。なお、実験条件の実施順序は被験者によって異なるようにした。

5. 実験結果と考察

実験後、未修正の折り返し翻訳文および被験者が修正したメッセージの折り返し翻訳文のそれぞれの精度の主観評価を Walker らの適合性評価¹⁵⁾ を用いて行った。Walker らの適合性評価は 5 段階評価であり、2 名以上の評価者により行われる。評価基準は、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)、である。本実験では、大学生 3 名を評価者とし、評価を行った。なお、各文

表 3 実験テキストの平均精度

Table 3 An average evaluated value in each experimental test set.

	未修正	修正後(精度表示なし)	修正後(精度表示あり)
平均精度	1.8	3.5	3.4
標準偏差	0.5	0.6	0.5
有意確率 ¹	0.000		

¹Friedman 検定を用いた。

表 4 未修正の精度と各条件下の精度との差の有意確率

Table 4 Significance probability in each experimental condition.

比較条件	有意確率 ¹
未修正・修正後(精度表示なし)	0.000*
未修正・修正後(精度表示あり)	0.000*
修正後(精度表示なし)・修正後(精度表示あり)	0.972

*: 有意差あり p<0.005 (多重比較を行うため, ボンフェローニの不等式を利用し, 有意水準 5%を比較対象となる組み合わせ数 10 で割った 0.5%を有意水準とし, 検定を行う)

¹ ウィルコクソンの符号付順位検定を用いた。

については, 3 人の評価者による評価の平均値を精度評価値として扱う。

未修正の折り返し翻訳の精度が良い場合, 被験者は修正作業を行う必要がない。本実験では, 修正の必要な精度の低い文 40 文と, 修正の不要な精度の高い文 20 文を用いた。そこで, 本稿では, 修正の必要な精度の低い文 40 文について分析を行う。

5.1 作成されたテキストの精度

未修正および修正後のテキストの平均精度を表 3 に示す。表 3 における値は, 分析対象である 40 文に対する評価の平均値である。評価の結果, 未修正のテキストの精度および各条件下における修正後のテキストの精度に有意差が見られた。また, これらのうち, どの組み合わせに有意な差があるかどうかを調べるために, ウィルコクソンの符号付順位検定により, 多重比較を行った。表 4 に未修正の精度と各条件下の精度との差の有意確率を示す(ボンフェローニの不等式より, 有意水準 0.5%において有意差があったものに*印を付加した)。多重比較の結果, 未修正のテキストの精度と各条件下における修正後の精度間に有意差が見られたが, 精度表示の有無における有意差は見られなかった。

5.2 修正放棄数

表 5 に, 各条件下における修正放棄数を示す。本研究の目的は, 精度表示を行うことによ

表 5 修正放棄数

Table 5 Number of the sentences given up in the experiment.

	精度表示なし	精度表示あり
修正放棄数	63	59

各条件下における母数は 240 文(40 文×6 名)である。

表 6 不正確判定数

Table 6 Number of the inaccurate judgments.

	精度表示なし	精度表示あり
不正確判定数	38	39

各条件下における母数は 240 文(40 文×6 名)である。

り, 不正確判定の減少効果の検証である。被験者による修正の放棄は, 折り返し翻訳文の精度は十分でないが, 5 分程度経過した際に行われた。そのため, 修正放棄された文については, 精度の判断が正確に行われており, 不正確判定ではないものとして扱う。

5.3 不正確判定数

本研究では, 精度評価に 5.1 節で述べた 5 段階評価を用いた。用いた指標において, 評価値が 3 (意味は何となく掴める) 未満である場合, 折り返し翻訳文から原文の意味を読み取ることができないことを意味する。被験者が修正完了したテキストのうち, 評価者による評価平均値が 3 未満である場合, 被験者の文意一致判定が不正確であると考えられる。そこで, 修正後の精度が 3 未満である文について, 不正確判定がなされたと判断することとし, 不正確判定数を確認した。

各実施条件における不正確判定数を表 6 に示す。表 6 より, 精度表示なしにおける不正確判定数は 38, 精度表示ありにおける不正確判定数は 39 となっており, 大きな差異は見られなかった。

精度表示を行った場合, 一定の精度に到達した段階で被験者が修正を終了した可能性がある。しかし, 精度評価は自動で行っているため, 実際の精度と一致しない場合がある。実際は精度が低くても表示された精度が良い場合, 表示された自動評価の影響を受け, 被験者が修正を終了していた可能性があると考えられる。

そこで, 不正確判定が発生したテキスト(人手評価値が 3 未満のテキスト)のうち, 「何となく意味はつかめる」以上の精度表示(自動評価値が 3 以上)がされていた場合の数を確認した。確認の結果, 表 6 に示した「精度表示あり」における不正確判定 39 回のうちの 10 回は, 自動評価が誤っていたことがわかった。これらの不正確判定については, 自動評

表 7 アンケート結果
Table 7 Results of questionnaire.

質問番号	質問	平均 (標準偏差)
1	折り返し翻訳に精度が表示されていたとき、精度を確認した。	4.4(0.7)
2	精度表示は、修正の必要性の有無についての判断の役に立った。	3.9(1.0)

5段階評価の評価値：1：強く同意しない，2：同意しない，3：どちらともいえない，4：同意する，5：強く同意する

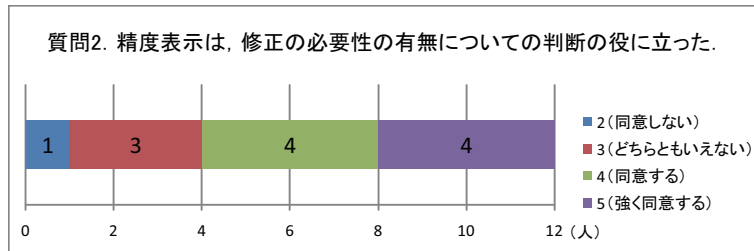


図 2 質問 2 の回答の分布
Fig. 2 Distribution of the answers of question No.2.

価の誤りが被験者に悪影響を与えたために発生した可能性がある。今後、自動評価の精度を向上し、自動評価の誤りを減少させることによって、不正確判定の数をより減少させることができる可能性があると考えられる。

5.4 アンケート結果

実験後に実施したアンケートの結果を表 7 に示す。表 7 の各質問に対する評価は、1：強く同意しない，2：同意しない，3：どちらともいえない，4：同意する，5：強く同意する，の 5 段階評価の平均値を示す。

質問 1 の評価平均は 4.4 となっており、被験者は精度表示を確認しながら修正作業を行っていたことがわかる。質問 2 の評価平均は 3.9 であった。質問 2 の回答の分布を図 2 に示す。「同意しない」と回答した被験者が 1 名、「どちらとも言えない」と回答した被験者が 3 名、「同意する」および「強く同意する」と回答した被験者が 8 名であった。「同意しない」と回答した被験者の自由記述においては、「精度が信用できない」という意見があった。「同意する」および「強く同意する」と回答した被験者の自由記述においては、「多くの場合は役に立つが、意味がわかる文でも精度が低い場合があった」という意見が多く、自動評価の精度に問題があると考えられる。

6. おわりに

本稿では、翻訳リペアにおけるユーザの不正確判定を防止するために、翻訳リペアにおける翻訳精度表示を提案した。翻訳精度表示を行うことによる、ユーザの不正確判定に対する効果を検証するために「%表示」と「5段階表示」を組み合わせた手法を用いてテキスト修正実験を行った。実験の結果、以下の知見を得た。

- (1) 本実験においては、精度表示を行うことによる不正確判定の減少効果は見られなかった。ただし、自動評価を用いたため、自動評価の失敗により不正確判定につながった場合もあると考えられる。今後自動評価の精度が向上することにより、不正確判定の減少効果につながる可能性があると考えられる。
- (2) アンケートの結果、精度表示は役に立つ可能性があるものの、自動精度評価の精度に問題があり、改善する必要があることがわかった。

今後は、自動精度評価の精度を改善し、より正確な精度の提供を行う。また、不正確判定を減少させるための手法についての検討を進める。

謝辞 なお本研究の一部は、日本学術振興会科学研究費 基盤研究 (B)(19300036) および基盤研究 (B)(22300044) の補助を受けた。

参 考 文 献

- 1) Aiken, M.: Multilingual Communication in Electronic Meetings, ACM SIG-GROUP, Bulletin, 23, 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, 33, 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, Lecture Notes in Computer Science 4560, pp.91-97 (2007).
- 5) 宮部真衣, 吉野孝, 重信智宏: 折返し翻訳を用いた翻訳リペアの効果, 電子情報通信学会論文誌, Vol.J-90-D-I, No.12, pp.3142-3150 (2007).
- 6) Miyabe, M., Yoshino, T. and Shigenobu, T.: Effects of Repair Support Agent for Accurate Multilingual Communication, Proceedings, Lecture Notes in Computer Science 5351, pp.1022-1027 (2008).
- 7) 宮部真衣, 吉野孝: 機械翻訳を介したチャットコミュニケーションにおける精度判定に基づく送信拒否の適用可能性, 情報処理学会論文誌, Vol.51, No. 3, pp.784-795 (2010).

- 8) Papineni, K., Roukos, S., Ward, T. and Zhu, W.: BLEU: a Method for Automatic Evaluation of Machine Translation, In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp.311-318 (2002).
- 9) NIST: Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics, Technical report, NIST (2002).
- 10) 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).
- 11) Denoual, E., 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).
- 12) 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).
- 13) Uchimoto, K., Hayashida, N., Ishida, T. and Isahara, H.: Automatic Rating of Machine Translatability, 10th Machine Translation Summit (MT Summit X), pp.235-242 (2005).
- 14) 宮部真衣, 吉野孝: 機械翻訳を介したテキストコミュニケーションのための翻訳精度表示の影響, マルチメディア, 分散, 協調とモバイル (DICOMO2010) シンポジウム, pp.117-125 (2010).
- 15) Kevin Walker, Moussa Bamba, David Miller, Xiaoyi Ma, Chris Cieri, and George Doddington, "Multiple-Translation Arabic (MTA) Part 1," Linguistic Data Consortium (LDC) catalog number LDC2003T18 and ISBN 1-58563-276-7.
- 16) Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237(2004).
- 17) 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).
- 18) 会話表現データベース, ATR 音声翻訳通信研究所, <http://www.atr-p.com/sdb.html>
- 19) Tomohiro Shigenobu: Evaluation and Usability of Back Translation for Intercultural Communication, Proceedings Lecture Notes in Computer Science 4560, pp.259-265(2007).
- 20) 金淵培, 江原暉将: 日英機械翻訳のための日本語長文自動短文分割と主語の補完, 情報処理学会論文誌, Vol.35, No.6, pp.1018-1028(1994).
- 21) 黒橋禎夫, 長尾眞: 長い日本語文における並列構造の推定, 情報処理学会論文誌, Vol.33, No.8, pp.1022-1031(1992).
- 22) 乾孝司, 乾健太郎: 複数のパーザを利用した統計的部分係り受け解析, 情報処理学会論文誌, Vol.42, No.12, pp.1234-1246(2001).
- 23) Toru Ishida, "Language Grid: An Infrastructure for Intercultural Collaboration," IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06), pp.96-100, 2006.
- 24) 高電社, <http://www.kodensha.jp/>