# 対訳コーパスを用いた翻訳品質自動評価法

安田圭志[†,††] 菅谷史昭[†,☆,☆☆] 竹澤寿幸[†]
山本誠一[†] 柳田益造[††]

翻訳品質の自動評価法を提案する．提案手法は，対訳コーパスから翻訳正解を補い，システムによる翻訳結果と翻訳正解とで，DP マッチングにより表層的類似度に基づく評価を行うものである．提案手法を ATR 音声翻訳通信研究所で研究開発された音声翻訳システム日英 ATR-MATRIX の言語翻訳部の評価に適用した結果について示し，主観評価との相関の観点から，その評価性能を検証する．提案手法の応用例として．提案手法での評価結果を用いて主観評価による訳質のランクの決定を自動化する方法について述べる．主観評価でまったく問題ない翻訳であると評価されるものと，それ以外の 2 クラス分けの判別では，81%と高い判別率が得られた．また，提案手法と主観評価を併用した場合，主観評価のコストを 30%削減可能であることが示された．

# Automatic Evaluation Method of Translation Quality Using Parallel Corpus

Keiji Yasuda,[†,††] Fumiaki Sugaya,[†,☆,☆☆] Toshiyuki Takezawa,[†]
Seiichi Yamamoto[†] and Masuzo Yanagida[††]

An automatic translation quality evaluation method is proposed. In the proposed method, a parallel corpus is used to query translation answer candidates. The translation output is evaluated by measuring the similarity between the translation output and translation answer candidates with DP matching. This method evaluates a language translation subsystem of the Japanese-to-English ATR-MATRIX speech translation system developed at ATR Interpreting Telecommunications Research Laboratories. Discriminant analysis is then carried out to decide translaton rank automatically. The discriminant ratio is 81% for 2-class discrimination between absolutely correct and less appropriate translations classified subjectively. Also discussed is the hybrid method of the proposed method and subjective evaluation. Using the hybrid method, achievable cost reduction of subjective evaluation is 30%.

## 1. はじめに

今日におけるインターネットの爆発的普及，経済のグローバル化等にともない，異なる言語を話す人同士がコミュニケーションする機会と必要性が増してきた．それにともない音声翻訳システムに関する研究が活発に行われている[1)~7)]．音声翻訳システムの研究において，システムの評価は不可欠である．たとえば，我々はこれまでに，ATR 音声翻訳通信研究所で研究開発された日英双方向音声翻訳システム ATR-MATRIX[8)] の翻訳結果を A，B，C，D の 4 ランクに評価者が主観で割り当て，翻訳ランクを決定する翻訳ランク評価法[9)] や，システムと人間能力との比較を通じてシステムの翻訳能力を評価する翻訳一対比較法を実施してきた[10)]．これらの評価手法は，主観による判定が必要であり，それに要するコストは少なくない．そのため，システムの研究・開発期間中に，頻繁に評価を実施することは難しい．低コストな自動評価法が利用できれば，より頻繁に評価を実施することができ，システム改善作業の効率化を図ることができる．

自動評価法としては，対訳テストセットを用いた DP マッチングによる評価手法[11),12)] が提案されている．この手法は，表層的単語の一致度に基づいて評価するため，対訳として用意されていない同義で別形式の表現の翻訳結果に対して一致度が小さくなるという問題がある．この問題を解決するため，同義の別形式

† ATR 音声言語コミュニケーション研究所
ATR Spoken Language Translation Research Laboratories
†† 同志社大学
Doshisha University
☆ 現在，KDDI 研究所
Presently with KDDI R&D Laboratories, Inc.
☆☆ 現在，神戸大学大学院
Presently with Graduate School of Kobe University

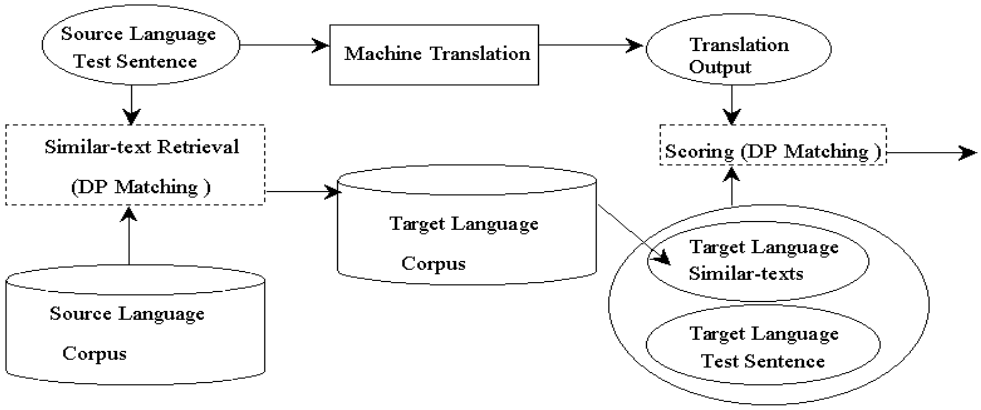


**図 1** 提案手法の構成
Fig. 1 The diagram of the proposed method.

の表現（パラフレーズ）を人手で大量に用意し，評価に利用する方法である「パラフレーズ利用型自動評価法」[13] が提案されており，この手法を用いた研究も行われている[14]．しかしながら，単語レベルの類似表現や語順の入替え，用いられる言い回しや基本文型までをも考慮にいれながら，表層レベルの多様な表現を網羅的に収集するのは容易な作業ではない．さらに，実際に運用される場面まで考慮すれば，表現や語順を組合せ的に考えれば解決するわけではなく，好んで使用される単語や語順，基本文型もあるため，そのような頻度情報についても考慮に入れた効率的なデータ収集法が研究課題となっている．

本論文では，既存の対訳コーパスからパラフレーズを自動的に収集し，評価に利用する「検索型自動評価法」を提案する．提案手法では，データ収集のコストが削減されるだけでなく，コーパスの特性を反映したパラフレーズを評価に利用することができる．そのため，利用場面の特性を反映したコーパスを利用すれば，利用場面で好んで使用される単語や語順の選好までをも考慮にいれたパラフレーズを，翻訳評価に利用することができる．

本論文は以下のように構成される．まず，2 章で検索型自動評価法についての説明を行う．3 章で検索型自動評価法による翻訳評価結果を示し，その評価性能について検討する．4 章では検索型自動評価法の応用例として，翻訳ランク評価を自動化する手法について述べる．5 章で全体をまとめる．

## 2. 検索型自動評価法

検索型自動評価法の処理の流れを**図 1** に示す．図中の，原言語コーパス（Source language corpus）と目的言語コーパス（Target language corpus），および，原言語テスト文（Source language test sentence）と目的言語テスト文（Target language test sentence）は対訳関係になっている．図 1 では，まず，DP マッチングにより，原言語側コーパスの中から，原言語側テスト文の類似文を検索する．類似度がある一定の閾値以上となるものを類似文と見なし，これを「類似原言語文」呼ぶ．また，ここでの閾値を「類似文検索閾値」と呼ぶ．類似文検索のための類似度 $\sigma_R$ を次式で定義する．

$$\sigma_R = \frac{T_R - S_R - I_R - D_R}{T_R} \tag{1}$$

ただし，$T_R$ は原言語コーパス内の各文における総語数，$S_R$ は原言語コーパス内の各文と原言語テスト文を DP マッチングにより比較したときの置換語数，$I_R$ は同様に比較したときの挿入語数，$D_R$ は同様に比較した場合の脱落語数である．

また翻訳評価のための類似度 $\sigma_E$ を，次式によって定義する☆．

$$\sigma_E = \frac{T_E - S_E - I_E - D_E}{T_E} \tag{2}$$

ただし，$T_E$ は正解目的言語文の総語数，$S_E$ は正解目的言語文とシステムによる翻訳結果を DP マッチングにより比較したときの置換語数，$I_E$ は同様に比較したときの挿入語数，$D_E$ は同様に比較した場合の脱落語数である．

次に，ここまでに得られた類似原言語文の目的言語側と，テスト文の目的言語側をあわせて「正解群」と

☆ $\sigma_E$ は，負の値をとりうるが，負の値となる場合は 0 としている．

する．最後に，言語翻訳結果と正解群の中の各文とで，DP マッチングにより類似度を求める．この結果として，正解群に含まれる文の数だけ類似度が求まるが，その最大類似度を「正解群類似度」（answer set similarity）とし，これを翻訳文の評価尺度とする．

式 (1) および式 (2) は音声認識の評価に用いられる音声認識率と同じ定義である．本論文では，音声認識率と同様に，置換，挿入，脱落誤りの影響をすべて等価な 1 としている．タスク，ドメインの性質によっては，異なる重みが適当となる場合も考えられるが，これについては，今後の検討課題とする．

なお，パラフレーズ利用型自動評価法においては，正解群の代わりに，人手で作成したパラフレーズを用いて同様の処理を行う．

## 3. 評価実験

検索型自動評価法による評価結果を示し，その評価性能について検討する．翻訳評価の対象は ATR-MATRIX の言語翻訳サブシステムである TDMT（Transfer Driven Machine Translation）[15] である．検索型自動評価法に用いた対訳コーパスは，ATR で構築された 618 会話（16110 文）からなるバイリンガル旅行対話データベース[16),17)] であり，テストセットはこのうちの 23 会話（330 文）である．この 23 会話は，言語翻訳部に対してオープンである．

自動評価法の評価性能は，翻訳ランクとの相関により検証する．翻訳ランクとしては，以下の基準で，5 人の評価者により決定された翻訳ランクを用いる．

**（A）**：訳文だけでまったく問題なし．
**（B）**：訳文は少し情報が欠けている．
**（C）**：訳文はかなり情報が欠けている．
**（D）**：訳文からは，情報が想像もできない．

相関を求めるにあたって，A ランクを 3，B ランクを 2，C ランクを 1，D ランクを 0 にそれぞれ数値化する．また，翻訳ランク評価法では，評価者ごとに評価結果が異なる場合があるため，各文について 5 人の評価者の平均（mean opinion score: MOS）を用いる．**図 2** に TDMT による翻訳結果の MOS で 0.25 ごとのヒストグラムを示す．図 2 のとおり，MOS が 3 となる翻訳結果，すなわち全評価者が A ランクと評価した翻訳結果が最も多く，全体の 30%以上を占めている．

### 3.1 類似文検索閾値について

**図 3** に，類似文検索閾値と検索により得られる正解群の平均文数の関係を示す．図中の横軸は類似文検索閾値を表しており，縦軸は文の数を表している．図中の○は，重複を許した場合ののべ文数，●は，重複を許さない場合の異なり文数を表している．原言語側（日本語）の類似文検索では，形態素単位での DP マッチングを行い，目的言語側（英語）でのスコアリングでは，単語単位での DP マッチングを行っている．図 3 より類似文検索閾値の減少にともない，追加される文の数が増加していることが分かる．**図 4** に類似文検索閾値と正解群類似度の関係を示す．図中の横軸は類似文検索閾値を表しており，縦軸は正解群類似度，または MOS との相関を表している．図中の○は MOS が 3 のテスト文の平均正解群類似度を，●は MOS が 1 以下となるテスト文の平均正解群類似度を表している．MOS が 3 の場合（図中の○）については，類似文検索閾値が 0.6 以下でほぼ同じ値となっている．このことは，類似文検索閾値を 0.6 未満にしても同義の目的言語文がほとんど得られないことを表している．また MOS が 1 以下の場合（図中の●）については，類似文検索閾値が 0.6 未満でも増加傾向にある．この見かけ上の増加は，類似文検索閾値の値を小さくしす

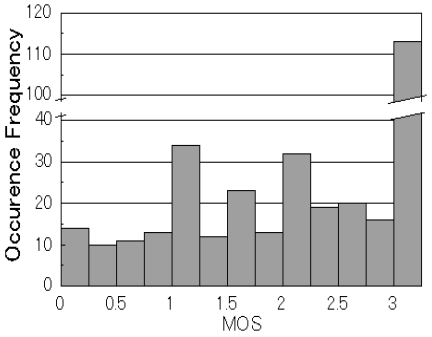


**図 2** MOS のヒストグラム
Fig. 2 Histgram of MOS.

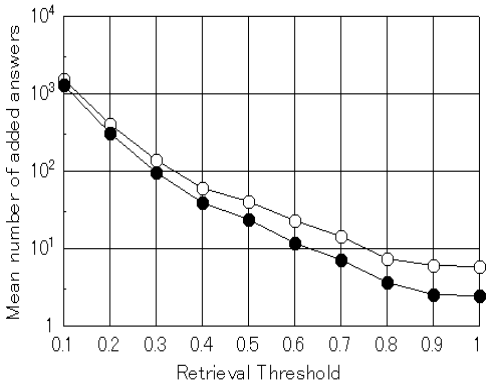


**図 3** 類似文検索閾値と検索により追加される文数の関係
Fig. 3 Relationship between the retrieval threshold and number of added sentences.

ぎると，異義の目的言語文が追加され，誤った翻訳文と部分的にマッチしてしまい，正解群類似度の値も大きくなることが原因と考えられる．最適な類似文検索閾値を，MOS との相関が最大となる閾値として決定する方法が考えられるが，図 2 をみると，MOS ごとのデータ数には偏りがあり，MOS が低いデータの数は少ない．このため，MOS が低いデータに対して異義の目的言語文が追加されるペナルティが，全データの相関には顕著に表れない．図 4 中の■が MOS との相関である．図 4 では，類似文検索閾値が 0.2 で相関が最大となっているが，類似文検索閾値が 0.1〜0.6 での相関の差は小さい．MOS の低いデータに対するペナルティを考慮すると，このような有意な差がでない MOS との相関による最適な類似文検索閾値の決定は困難である．

そこで，最適な類似文検索閾値は，以下に定義する評価関数 $F_{opt}$ により決定する．$F_{opt}$ が最大となる閾値を最適な類似文検索閾値とする．

$$F_{opt} = m_{MOS3} - m_{MOS1} \quad (3)$$

$m_{MOS3}$ は，MOS が 3 のテスト文の正解群類似度の平均を，$m_{MOS1}$ は MOS が 1 以下のテスト文の正解群類似度の平均をそれぞれ表す．$F_{opt}$ は，それぞれの MOS の正解群類似度の平均を用いていることから，データの偏りの影響はない．図 4 中の□が，$F_{opt}$ の値である．図 4 で，$F_{opt}$ の値は，類似文検索閾値が 0.6 でピークとなっている．これらの結果から，類似文検索閾値の値としては 0.6 が最適であると考えられる．

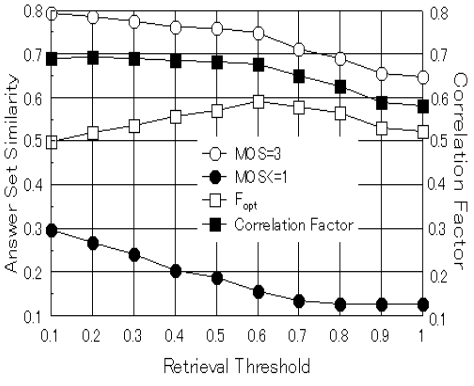


**図 4** 類似文検索閾値と正解群類似度の関係
Fig. 4 Relationship between the retrieval threshold and answer set similarity.

### 3.2 検索型自動評価法による翻訳評価結果

**表 1** は，類似文検索により追加される正解群の一例である．表中の “/” は，日本語における語境界を表している．類似原言語文の追加によって，目的言語側でも異なる言い回しではあるが，同義の文が得られていることが分かる．

**図 5** と**図 6** に検索型自動評価法による TDMT の評価結果を示す．図 5 中の横軸は翻訳ランクを，縦軸は正解群類似度を表している．ここでは，5 人の評価者によって決定された各文の翻訳ランクの中央値（Median）を，各文の翻訳ランクとしている．また，横軸ラベルの括弧内の数字は，各ランクの文数である．図 6 中の横軸は MOS を，縦軸は正解群類似度を表している．図 5 と図 6 中の各点は各翻訳ランクおよび各 MOS における正解群類似度の平均を，エラーバーは標準偏差を表している．図 5 と図 6 では，翻訳ランクおよび MOS が高くなればなるほど，正解群類似度の平均も大きくなる傾向がみられる．図 5 の翻訳ランク C ランク，D ランクについて，逆の結果となっているが，これは，提案手法による C ランク以下の評価が難しいということを示唆している．

### 3.3 パラフレーズ利用型自動評価法との評価性能の比較

パラフレーズ利用型自動評価法は，正解翻訳に対するカバレージを上げるうえで有効な手法であるが，1 章で述べたように，パラフレーズ作成に多くのコストがかかるという問題がある．これに対し検索型自動評価法では人手がまったく介入しないため，コスト面での有効性は明らかであるが，ここでは，両手法を評価性能の観点からの比較，検討を行う．

パラフレーズ利用型自動評価法に用いるパラフレー

**表 1** 正解群の一例
Table 1 Examples of the answer set.

| Source language test sentence | Target language test sentence |
|---|---|
| はい/分かり/まし/た/お/調べ/し/ます/ので/少々/お/待ち/ください | All right. Please hold the line and I will check. |

| Source language similar-texts | Target language similar-texts |
|---|---|
| かしこまりました/お/調べ/いたし/ます/ので/少々/お/待ち/ください | Okay, let me check. Just a moment please. |
| はい/お/調べ/し/ます/少々/お/待ち/ください/ませ | Okay, could you wait for a moment while I check. |
| 分かり/まし/た/確認/し/ます/ので/少々/お/待ち/ください | Okay, I 'll check for you please hold on a moment. |
| お/調べ/いたし/ます/ので/少々/お/待ち/ください | One moment please. I 'll check on availability. |
| ただいま/お/調べ/し/ます/ので/少々/お/待ち/ください/ませ | Could you hold on a minute while I check please. |

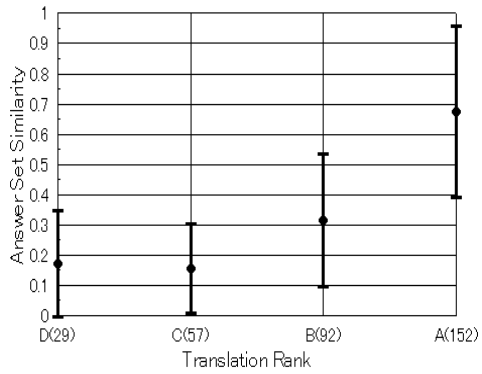


**図 5** 翻訳ランクと正解群類似度の関係
Fig. 5 Relationship between translation rank and answer set similarity.

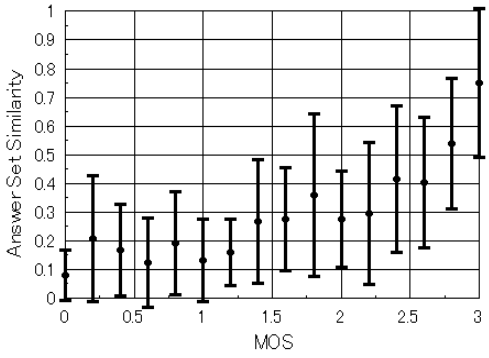


**図 6** MOS と正解群類似度の関係
Fig. 6 Relationship between MOS and answer set similarity.

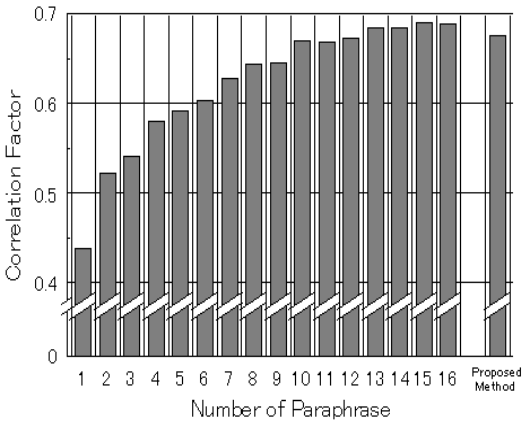


**図 7** 検索型自動評価法とパラフレーズ利用型自動評価法の評価性能の比較
Fig. 7 Performance comparison between the proposed method and a conventional method.

**表 2** 相関行列
Table 2 Correlation matrix.

| | MOS | Evaluator 1 | Evaluator 2 | Evaluator 3 | Evaluator 4 | Evaluator 5 |
|---|---|---|---|---|---|---|
| MOS | 1.000 | | | | | |
| Evaluator 1 | 0.913 | 1.000 | | | | |
| Evaluator 2 | 0.875 | *0.749* | 1.000 | | | |
| Evaluator 3 | 0.877 | *0.770* | *0.719* | 1.000 | | |
| Evaluator 4 | 0.931 | *0.836* | *0.758* | *0.772* | 1.000 | |
| Evaluator 5 | 0.924 | *0.804* | *0.742* | *0.751* | *0.824* | 1.000 |
| #of paraphrase = 1 | 0.439 | **0.377** | **0.368** | **0.376** | **0.411** | **0.438** |
| #of paraphrase = 16 | 0.690 | **0.617** | **0.598** | **0.598** | **0.633** | **0.661** |
| Proposed method | 0.677 | **0.589** | **0.609** | **0.592** | **0.596** | **0.662** |

ズデータは，日本語が分かる 5 人の英語ネイティブ話者が，各テスト文につき，3 文のパラフレーズをすることにより収集した．5 人が個別にパラフレーズを行うため．パラフレーズの結果が，同じ文になる場合もあるが，その数は少なく，1 文の原言語文につき，平均で 14.4 文の異なり目的言語文が収集されている．パラフレーズ利用型自動評価法においては，ここで収集したデータに加え，対訳テストセットの英語側をも用いるため，正解数が最大で 16 文となる．

**図 7** に両手法による評価結果と MOS との相関を示す．縦軸は MOS との相関係数を表しており，横軸はパラフレーズ利用型自動評価法で用いたパラフレーズ数および，検索型自動評価法を表している．図 7 より，パラフレーズ利用型自動評価法では，パラフレーズ数が増えれば増えるほど MOS との相関が高くなっていることが分かる．また，検索型自動評価法においては，パラフレーズ利用型自動評価法でパラフレーズ数を 16 文とした場合よりは多少劣る（MOS との相関係数で 0.013）ものの，ほぼ同等の評価性能が得られている．**表 2** は各評価者ごとのランク評価の結果と，各自動評価法による評価結果の相関行列である．ここでの自動評価法は，検索型自動評価法と，パラフレーズ利用型自動評価法でパラフレーズ数を 1 および 16 とした場合である．表中の斜体は，評価者間の相関を表しており，表中の太字が各自動評価法と各評価者との相関を表している．表 2 において，評価者間の相関をみると，0.719〜0.836 となっており，評価者 2 と評価者 3 との相関が最も低く（表中の下線），評価者 1 と評価者 4 との相関が最も高い（表中の二重下線）．検索型自動評価法と各評価者との相関は 0.589〜0.662 であり，評価者間の相関には及ばないが，パラフレーズ利用型自動評価法でパラフレーズ数を 1 とした場合と各評価者との相関（0.368〜0.438）よりは格段に高く，また，パラフレーズ数を 16 とした場合と各評価者との相関（0.598〜0.661）とほぼ同等になっている．

**表 3** 判別分析の結果
Table 3 Result of discriminant analysis.

| | A/BCD | AB/CD | ABC/D | A/B/C/D |
|---|---|---|---|---|
| Discriminant Ratio | 0.81 | 0.72 | 0.62 | 0.57 |

## 4. 検索型自動評価法の応用

これまでに述べたように，検索型自動評価法は，主観評価と相関の高い評価を行えることが分かった．そこで本章では，検索型自動評価法を用いた翻訳ランクの自動判別について検討する．なお，本章で用いたテストセットは 3 章で述べたものと同様の，23 会話（330 文）からなるテストセットである．

### 4.1 判別分析によるランク決定

判別は，各クラスの正解群類似度の平均を求め，最近傍則に従って行う．

以下に判別率（Discriminant ratio）$D$ を定義する．

$$D = n_{correct}/n_{total} \quad (4)$$

ただし，$n_{correct}$ は正しく判別された文の数，$n_{total}$ は，全体の文の数である．

判別分析に用いた翻訳ランクは，5 人の評価者によって決定された各文の翻訳ランクの中央値である．また，判別分析に用いた各クラスの平均は，テストセット 330 文から求めている．**表 3** に判別分析の結果を示す．

ここでは，2 クラス分けの判別と，4 クラス分けの判別を行っている．2 クラス分けの判別では，A ランクと B，C，D ランクの 2 クラス分け（表中の A/BCD），A，B ランクと C，D ランクの 2 クラス分け（表中の AB/CD），A，B，C ランクと D ランクの 2 クラス分けを行った（表中の ABC/D）．4 クラス分けについては，翻訳ランク評価法の各ランクを，そのままクラスとしている（表中の A/B/C/D）．表 3 では，A ランクと B，C，D ランクの 2 クラス分けの判別率が特に高く，0.81 となっている．また，4 クラス分けの場合は，判別率が 0.57 と低い．

**表 4** は 5 人の各評価者による評価結果と，全評価者による評価結果の中央値との一致率である．一致率 $A$ を次式で定義する．

$$A = n_{accord}/n_{total} \quad (5)$$

ただし，$n_{accord}$ は各評価者による評価結果と，全評価者の評価結果の中央値が一致した文の数である．4 クラス分けの一致率を計算する場合は，翻訳ランクそのものの値を比較するが，2 クラス分けの場合については，4 クラスから 2 クラスに情報を落としたうえで一致率を計算している．表 4 のラベルについては，表 3 と同様の表記方法である．表 4 において，4 クラス分けについては，各評価者と中央値の一致率も 2 クラス分けの場合と比較して，0.69〜0.79 と低くなっており，人が評価する場合でも，評価のゆれが大きいことを表している．

**表 4** 各評価者の評価結果と中央値との一致率
Table 4 Accordance ratio between median and the evaluation result by each evaluator.

| | A/BCD | AB/CD | ABC/D | A/B/C/D |
|---|---|---|---|---|
| Evaluator 1 | 0.92 | 0.92 | 0.95 | 0.79 |
| Evaluator 2 | 0.92 | 0.91 | 0.92 | 0.79 |
| Evaluator 3 | 0.92 | 0.87 | 0.96 | 0.75 |
| Evaluator 4 | 0.93 | 0.92 | 0.94 | 0.79 |
| Evaluator 5 | 0.94 | 0.89 | 0.81 | 0.69 |

**表 5** 2 クラス判別の 4 つの確率
Table 5 4 kinds of probability for 2-class discrimination.

| | | State | |
|---|---|---|---|
| | | *class 1* | *class 2* |
| Automatic Discrimination | *CLASS 1* | *P(CLASS 1 \| class 1)* | *P(CLASS 1 \| class 2)* |
| | *CLASS 2* | *P(CLASS 2 \| class 1)* | *P(CLASS 2 \| class 2)* |

### 4.2 検索型自動評価法と翻訳ランク評価法との併用方法

4.1 節では，A ランクと B，C，D ランクの 2 クラス分けは，判別率が 0.81 と，高い精度での自動判別が可能であることが示された．4 クラス分けについては，各評価者の評価結果とそれらの中央値との一致率も低いものの，自動判別による判別率が 0.57 と低い．実用を考えた場合には，より高い精度が要求される状況が考えられる．そこで，検索型自動評価法と翻訳ランク評価法とを併用し，低コストで高い精度の 4 クラス分けを行う方法について検討する．

検索型自動評価法と翻訳ランク評価法との併用では，A ランクの判別を自動で行い，A ランク以外に属すると自動判別されたものだけについて，翻訳ランク評価法により評価を行う．以降，A ランクをクラス 1，B，C，D ランクをクラス 2 と呼ぶ．

クラス 1 とクラス 2 の判別においては，5 人の評価者によって決定された各文の翻訳ランクの中央値が A ランクである場合（$class\,1$）と，B，C，D ランクである場合（$class\,2$），それをクラス 1 であると自動判別する場合（$CLASS\,1$）と，クラス 2 であると自動判別する場合（$CLASS\,2$）の 4 つの組合せがあり，**表 5** に示す 4 種類の確率が定義できる．表 5 において，$P(CLASS\,1\,|\,class\,1)$ はクラス 1 をクラス 1 として正しく受理する確率（正解受理率：Correct acceptance ratio），$P(CLASS\,1\,|\,class\,2)$ はクラス

2 をクラス 1 として誤って受理する確率（誤り受理率：False acceptance ratio），$P(CLASS\,2\,|\,class\,1)$ はクラス 1 をクラス 2 として棄却する確率（棄却誤り率：False rejection ratio），$P(CLASS\,2\,|\,class\,2)$ はクラス 2 をクラス 2 として棄却する確率（棄却正解率：Correct rejection ratio）である．

### 4.3 コストと誤りの定義

検索型自動評価法と翻訳ランク評価法との併用を考えた場合，棄却誤り率は誤りの指標とはならない．なぜなら，クラス 2 と自動判別された結果については再度翻訳ランク評価法で評価されるからである．しかしながら，棄却誤りは，翻訳ランク評価の評価コストを増やす要因となる．そこで，検索型自動評価法と翻訳ランク評価法との併用の観点からのコスト削減率と誤り率を定義する．コスト削減率 $P(CLASS\,1)$ は，自動判定されるテスト文の，テストセット全体に対する割合として次式で定義する．

$$P(CLASS\,1) = P(CLASS\,1\,|\,class\,1) \times P(class\,1) + P(CLASS\,1\,|\,class\,2) \times P(class\,2) \quad (6)$$

ただし，$P(class\,1)$ は全テストセットのうちクラス 1 に属する文の割合で，$P(class\,2)$ は全テストセットのうちクラス 2 に属する文の割合である．今回の評価対象である TDMT では，$P(class\,1)$ が 0.46，$P(class\,2)$ が 0.54 である．

式 (6) の右辺第 1 項の $P(CLASS\,1\,|\,class\,1) \times P(class\,1)$ は，クラス 1 をクラス 1 として正しく受理する文数の，テストセット全体に対する割合である．また，右辺第 2 項の $P(CLASS\,1\,|\,class\,2) \times P(class\,2)$ は，クラス 2 を誤ってクラス 1 として受理してしまう文数の，テストセット全体に対する割合である．

誤り率 $E$ については，クラス 1 と自動判定された中に含まれる誤りの割合として次式で定義する．

$$E = \frac{P(CLASS\,1\,|\,class\,2) \times P(class\,2)}{P(CLASS\,1)} \quad (7)$$

### 4.4 TDMT への適用結果

**図 8** は，判別する際にクラスの境界とする正解群類似度（判別閾値）と，棄却誤り率および誤り受理率の関係である．横軸は，判別閾値を表しており，縦軸は，誤り受理率，または棄却誤り率の値を表している．○は棄却誤り率を表しており，●は誤り受理率を表している．図 8 より，判別閾値を○と●の交点となる 0.4 として判別を行うと，誤り受理率および棄却誤り率は 0.2 となることが分かる．**図 9** は ROC 曲線（receiver-operating-characteristic curve）で，図 8 から求めた正解受理率と誤り受理率との関係である．横軸は誤り受理率，縦軸は正解受理率である．図中の各点は，図 8 における判別閾値を 0 から 1.0 まで，0.1 きざみで変化させた場合の結果である．

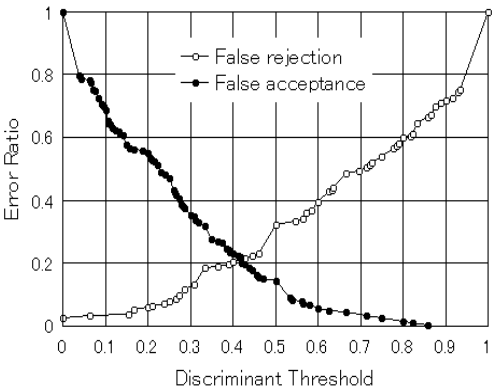


**図 8** 判別閾値と棄却誤り率および誤り受理率の関係
Fig. 8 Relationship between discriminant threshold and ratio for false rejection and false acceptance.

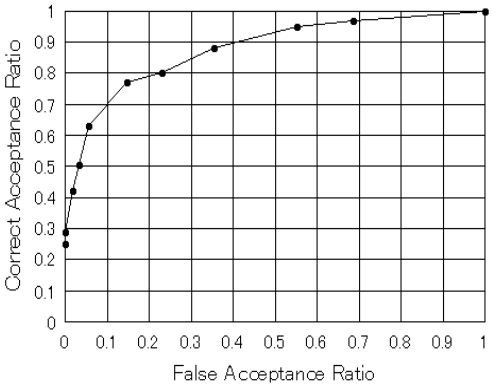


**図 9** ROC 曲線
Fig. 9 ROC curve.

**図 10** に，削減される評価コストと誤り率の関係を示す．横軸が式 (7) で定義した誤り率で，縦軸が式 (6) で定義したコスト削減率である．

各評価者による評価結果と全評価者による評価結果の中央値とのずれをみるため，各評価者による評価結果と中央値とを比較した場合の誤り率 $E_e$ を次式に定義する．

$$E_e = \frac{P(CLASS\,1_e\,|\,class\,2) \times P(class\,2)}{P(CLASS\,1_e)} \quad (8)$$

ここで，$CLASS\,1_e$ は各評価者による評価結果がクラス 1 である場合を表している．

**図 11** に評価者ごとの誤り率 $E_e$ を示す．図 11 をみると，各評価者による評価結果と中央値を比較した場合，誤り率は 1%〜14%となっており，5 評価者の平均で，9%の誤りが生じている（図 11 中の破線）．A ランクを自動判定する場合も，9%の誤りを許容した

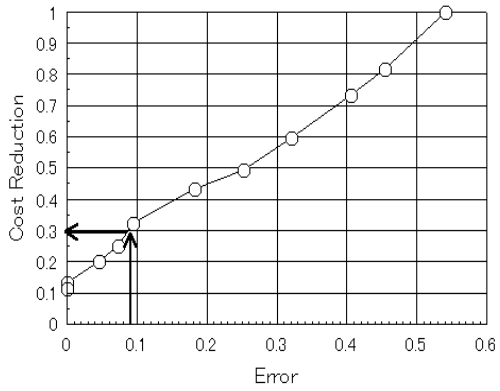


**図 10** 削減されるコストと誤り率の関係
Fig. 10 Relationship between cost reduction ratio and error ratio.

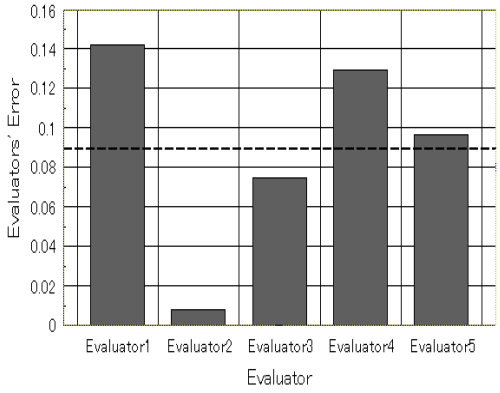


**図 11** 各評価者の誤り率
Fig. 11 Error ratio of each evaluator.

とすると，図 10 より，翻訳ランク評価の評価コストを約 30%削減可能であることが分かる．

## 5. むすび

翻訳品質の自動評価法として，検索型自動評価法を提案した．また，主観評価との相関に着目し，提案手法の評価性能の検証を行った．この結果，提案手法によって，人手でパラフレーズを作成し自動評価に利用するパレフレーズ利用型自動評価法とほぼ同等の評価性能が得られることが示された．

検索型自動評価法の応用例としては，従来の主観評価である翻訳ランク評価法の自動化法について述べた．また，より少ない誤りで，主観評価の評価コストを削減することを指向して，検索型自動評価法と翻訳ランク評価法の併用方法を提案し，それによって削減される評価コストを示した．現状の TDMT の評価への適用では，人による評価のゆれと同程度の誤りを許容した場合，評価コストを約 30%削減可能であることが示された．式 (1) および式 (2) における置換，挿入，脱落に対する最適な重みについては今後の検討課題とする．

**謝辞** 本研究を行ううえで，貴重なご指導をいただいた ATR 音声言語コミュニケーション研究所菊井玄一郎第二研究室室長に心より感謝いたします．本研究の一部は同志社大学学術フロンティア事業の援助を受けた．

## 参考文献

1) Hatazaki, K., Noguchi, J., Okumura, A., Yoshida, K. and Watanabe, T.: Intertalker: An Experimental Automatic Interpretation System using Conceptual Representation, *Proc. ICSLP*, pp.393–396 (1992).
2) 森元 逞，田代敏久，竹澤寿幸，永田昌明，谷戸文廣，浦谷則好，鈴木雅実，菊井玄一郎：音声翻訳実験システム（ASURA）のシステム構成と性能評価，情報処理学会論文誌，Vol.37, No.9, pp.1726–1735 (1996).
3) Rayner, M., Bretan, I., Carter, D., Collins, M., Digalakis, V., Gambäc, B., Kaja, J., Karlgren, J., Lyberg, B., Pulman, S., Price, P. and Samuelsson, C.: Spoken Language Translation with Mid-90's Technology: A Case Study, *Proc. EUROSPEECH*, pp.1299–1302 (1993).
4) Roe, D.B., Moreno, P.J., Sproat, R.W., Pereira, F.C.N., Riley, M.D. and Macarrón, A.: A Spoken Language Translator for Restricted-domain Context-free Languages, *Speech Communication*, Vol.11, pp.311–319 (1992).
5) Suzuki, M., Inoue, N., Yato, F., Takeda, K. and Yamamoto, S.: A Prototype of a Japanese-Korean Realtime Speech Translation System, *Proc. EUROSPEECH*, pp.1951–1954 (1995).
6) Bub, T., Wahlster, W. and Waibel, A.: Verbmobil: The Combination of Deep and Shallow Processing for Spotaneous Speech Translaton, *Proc. ICASSP*, pp.71–74 (1997).
7) Lavie, A., Waibel, A., Levin, L., Finke, M., Gates, D., Gavaldà, M., Zeppenfeld, T. and Zhan, P.: JANUS-: Speech-to-speech Translation in Multiple Language, *Proc. ICASSP*, pp.99–102 (1997).
8) Takezawa, T., Morimoto, T., Sagisaka, Y., Campbell, N., Iida, H., Sugaya, F., Yokoo, A. and Yamamoto, S.: A Japanese-to-English Speech Translation System: ATR-MATRIX, *Proc. ICSLP*, pp.2779–2782 (1998).
9) Sumita, E., Yamada, S., Yamamoto, K., Paul, M., Kashioka, H., Ishilawa, K. and Shirai, S.: Solutions to Problems Inherent in Spoken lan-

guage Translation: The ATR-MATRIX Approach, *Proc. MT Summit*, pp.229–235 (1999).
10) Sugaya, F., Takezawa, T., Yokoo, A., Sagisaka, Y. and Yamamoto, S.: Evaluation of the ATR-MATRIX Speech Translation System with Pair Comparison Method Between the System and Humans, *Proc. ICSLP*, pp.1105–1108 (2000).
11) Su, K.-Y., Wu, M.-W. and Chang, J.-S.: A New Quantitative Quality Measure for Machine Translation Systems, *Proc. COLING*, pp.433–439 (1992).
12) Takezawa, T., Sugaya, F., Yokoo, A. and Yamamoto, S.: A New Evaluation Method for Speech Translation Systems and a Case Study on ATR-MATRIX from Japanese to English, *Proc. MT Summit*, pp.299–307 (1999).
13) Thompson, H.S.: Automatic Evaluation of Translation Quality: Outline of Methodology and Report on Pilot Experiment, *Proc. Evaluators' Forum*, pp.215–223 (1991).
14) Akiba, Y., Imamura, K. and Sumita, E.: Using Multiple Edit Distance to Automatically Rank Machine Translation Output, *Proc. MT Summit VIII*, pp.15–20 (2001).
15) 古瀬　蔵，山本和英，山田節夫：構成素境界解析を用いた多言語話し言葉翻訳，自然言語処理，Vol.6, No.5, pp.63–91 (1999).
16) Morimoto, T., Uratani, N., Takezawa, T., Furuse, O., Sobashima, Y., Iida, H., Nakamura, A., Sagisaka, Y., Higuchi, N. and Yamazaki, Y.: A Speech and Language Database for Speech Translation Research, *Proc. ICSLP*, pp.1791–1794 (1994).
17) Takezawa, T.: Buildng a Biligual Travel Conversation for Speech Translation Research, *Proc. 2nd International Workshop on East-Asian Language-Resources and Evaluation — Oriental COCOSDA Workshop '99*, pp.17–20 (1999).

(平成 13 年 11 月 15 日受付)
(平成 14 年 4 月 16 日採録)

**安田　圭志**（学生会員）

平成 11 年同志社大学工学部知識工学科中退．平成 13 年同大学院修士課程修了．現在，同大学院博士後期課程在学中．ATR 音声言語コミュニケーション研究所研修研究員．日本バイオインフォマティクス学会会員．

**菅谷　史昭**（正会員）

昭和 57 年東北大学工学部通信工学科卒業．昭和 59 年同大学院修士課程修了．同年 KDD（株）入社．平成 3 年度学術奨励賞受賞．平成 9 年 ATR 音声翻訳通信研究所に出向．音声翻訳システム，言語翻訳評価の研究に従事．平成 13 年 4 月より神戸大学大学院在学中．平成 14 年 4 月より（株）KDDI 研究所に勤務．電子情報通信学会，日本音響学会各会員．

**竹澤　寿幸**（正会員）

昭和 59 年早稲田大学理工学部電気工学科卒業．平成元年同大学院博士後期課程修了．工学博士．昭和 62 年より同大学情報科学研究教育センター助手．平成元年より ATR 自動翻訳電話研究所に勤務．現在 ATR 音声言語コミュニケーション研究所主任研究員．音声翻訳システム，音声言語情報処理の研究に従事．電子情報通信学会，人工知能学会，日本音響学会，言語処理学会各会員．

**山本　誠一**

昭和 47 年大阪大学工学部電子工学科卒業．昭和 49 年同大学院修士（制御）課程修了．同年国際電信電話入社．以来，ディジタルファクシミリ，エコーキャンセラ，音声符号化，音声合成，音声認識，自然言語処理の研究に従事．平成 9 年 ATR 音声翻訳通信研究所に出向．現在，ATR 音声言語コミュニケーション研究所所長．昭和 56 年度学術奨励賞，日本音響学会第 3 回技術開発賞，日本音響学会第 5 回技術開発賞各受賞．著書「エコーキャンセラ技術」（共著）等．日本音響学会理事・関西副支部長，電子情報通信学会・情報・システムソサイエティ副会長，IEEE 会員．神戸大学大学院自然科学研究科客員教授．工学博士．

**柳田 益造**（正会員）

昭和 44 年大阪大学工学部電子工学科卒業．昭和 46 年同大学院修士課程修了．同年 NHK 入局．昭和 47 年大阪大学産業研究所研究生．昭和 53 年同大学院博士課程修了．工学博士．同年大阪大学産業研究所助手，昭和 53 年〜54 年オランダ国立 Groningen 大学音声研究所客員研究員，昭和 62 年大阪大学産業研究所助教授．同年郵政省電波研究所音声研究室長，平成元年同（通信総合研究所と改称）関西先端研究センターを経て，平成 6 年より同志社大学工学部知識工学科教授．音響信号処理，音声言語情報処理ならびに音楽知覚・情報処理の研究に従事．著書「ファジイ科学」（分担執筆),「信号処理」（分担執筆）．電子情報通信学会編集委員，日本音響学会理事・関西支部長，日本認知科学会，言語処理学会，音楽知覚認知学会，IEEE，Acout. Soc. Am.，Am. Psychol. Soc. 各会員．