音声言語情報処理 34－37
(2000. 12. 22)

対訳コーパスを用いた表層的類似度に基づく翻訳能力自動評価法

安田 圭志†‡ 菅谷 史昭† 竹澤 寿幸† 山本 誠一† 柳田 益造‡

†ATR 音声言語通信研究所 〒619-0288 京都府相楽郡精華町光台 2 丁目 2 番地 2 号
E-mail: keiji.yasuda@slt.atr.co.jp
‡同志社大学大学院工学研究科 〒610-0394 京田辺市多々羅都谷 1-3

あらまし 翻訳システムの自動評価手法を提案する．この提案手法は，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なうものである．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す．次に判別分析を行い，提案手法により得られる評価結果と従来の人手による主観評価での結果を比較している．この結果，主観評価で全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，83.5%と高い判別率が得られた．最後に，提案手法を言語翻訳部に音声認識を加えた音声翻訳システム全体の評価に適用する場合の問題点について述べる．
キーワード 音声翻訳システム，翻訳システム，翻訳自動評価，対訳コーパス，DP マッチング

An automatic evaluation method of translation capability by DP matching using similar expressions queried from a parallel corpus

Keiji YASUDA†‡, Fumiaki SUGAYA†, Toshiyuki TAKEZAWA†, Seiichi YAMAMOTO†, and Masuzo YANAGIDA‡

†ATR Spoken Language Translation Research Laboratories
2-2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto,619-0288,JAPAN
E-mail: keiji.yasuda@slt.atr.co.jp
‡Graduate School of Engineering, Doshisha University
1-3, Tatara-miyakodani, Kyotanabe, Kyoto, 610-0394, JAPAN

Abstract Proposed is an automatic evaluation method for translation system. A parallel corpus is used to query similar expressions of translation answers. Translation output is evaluated by measuring similarity between translation output and similar expressions of translation answers with DP matching. Evaluation by this method is conducted on the language translation subsystem of the Japanese-to-English ATR-MATRIX speech translation system developed at ATR Interpreting Telecommunications Research Laboratories. Discriminant analysis is carried out to analyze relationship between evaluation results of the proposed method and subjective evaluation. The experimental results show the effectiveness of the proposed method. Discriminat ratio is 83.5% in 2 class discrimination between absolutely correct and less appropriate translations so classified by human labelers. Also discussed are issues of the proposed method when applied to evaluate outputs of speech translation systems which make recognition errors.
Key words Speech translation system, Translation system, Automatic evaluation of translation, Parallel corpus, DP matching

1. はじめに

翻訳システムの性能改善の効率化や主観評価のコストを削減するためには，言語翻訳の自動評価技術が必要である．われわれは，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[1]の評価を通じて翻訳システムの評価方法について研究を進めてきた．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[2]や，システムと人間能力との比較を通じたシステムの翻訳能力評価を実施してきた[3]．しかしながら，何れの方法も，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．システムの性能改善の効率化には客観評価が必要となってきた．

客観評価手法としては，DP マッチングによる評価手法が提案されている[4][5]が，正解表現として登録されていない未登録の正解に対して DP マッチングに基づく類似度が小さくなるという問題があった．本論文では，DP マッチングをベースとしながら，未登録の正解を対訳コーパスで補う方法を提案する．次に，提案手法を日英 ATR-MATRIX の評価に適用し，この評価結果と翻訳ランクを用いて判別分析を行った結果について述べる．

2 章で正解文追加類似度計算法についての説明を行なう．3 章では，まず，正解文追加類似度計算を日英 ATR-MATRIX の言語翻訳部の翻訳評価に適用した結果を示し，次に，正解文追加類似度計算法を言語翻訳部の入力側に音声認識部を加えた音声翻訳統合評価に適用し，音声認識誤りが含まれる場合の問題点について述べる．4 章で全体をまとめる．

2. 正解文追加類似度計算法

従来の翻訳評価のための類似度(Similarity)は，以下の様に定義されている．

$$Similarity = \frac{Total - Sub - Ins - Del}{Total} \quad (1)$$

ここで，*Total* は正解翻訳文の総語数，*Sub* は正解翻訳文と翻訳システムからの翻訳出力を比較した時の置換語数，*Ins* は同様に比較した時の挿入語数，*Del* は同様に比較した場合の脱落語数である．

従来の DP マッチングによる類似度の問題点は，ある 1 つの原言語テスト文に対して，正解翻訳結果を 1 つだけしか用意していない点であった．正解文追加類似度計算法では，従来の DP マッチングによる類似度と同様に，翻訳結果と正解文との表層的な単語一致度に注目して評価しているが，従来の DP マッチングの類似度の問題点を解決するため，1 つの原言語テスト文に対して，複数の翻訳結果を対訳コーパスから収集している．1 つの原言語文に対して，複数の翻訳結果を持つパラフレーズコーパスが利用可能であれば最善であるが，現時点では，そのようなパラフレーズコーパスは利用可能ではない．

正解文追加類似度計算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），及び，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する．以降，ここで得られた類似文を類似原言語文，閾値を類似文検索閾値と呼ぶ．類似文検索の際の類似度は，式(1)で定義した類似度において，正解翻訳文を原言語テスト文に，翻訳システムからの翻訳出力を原言語コーパス内の文に置き換えて計算した類似度である．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 マッチングにより言語翻訳結果のスコアリングを行ない，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．

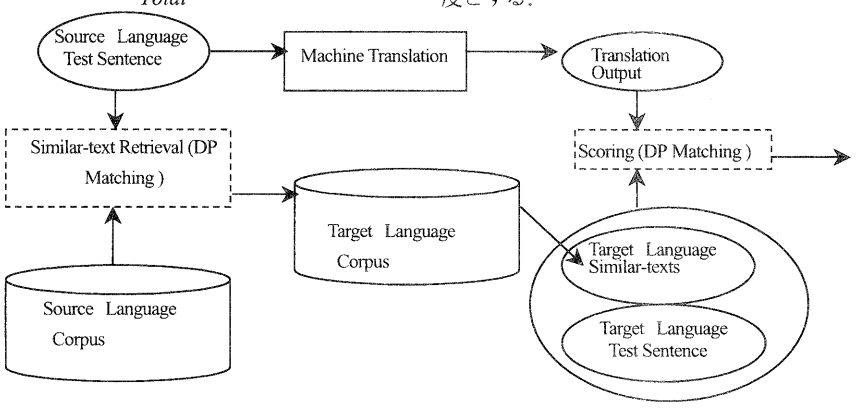

図 1　正解文追加類似度計算法の処理の流れ

3. 正解文追加類似度の評価結果

本章では，まず，従来の人手による翻訳ランク評価法について簡単に説明し，正解文追加類似度計算法をATR-MATIRX の言語翻訳サブシステムである TDMT(Transfer Driven Machine Translation)の評価に適用した結果について述べる．次に，これらの評価結果を用いた判別分析の結果を示す．最後に，言語翻訳部の入力側に音声認識部を加えた音声翻訳統合評価に適用した結果について示し，提案手法を音声翻訳システムの評価に適用する場合の問題点について述べる．

TDMT 単体の評価では，テストセットを TDMT へテキスト入力してえられた翻訳結果の評価を行ない，音声翻訳統合評価では，テストセットを音声認識部に音声入力し，認識結果を TDMT で翻訳した結果の評価を行なう．音声認識正解率は，88.5%である．

ここで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であり，テストセットはこの内の 23 会話（330 文）である．この 23 会話は，音声認識部，言語翻訳部に対してオープンである．

3.1 翻訳ランク評価

従来の人手による評価手法である翻訳ランク評価法では，評価者は主観により次の基準でランク付けを行っている．

(A) 完全訳: 訳文だけで全く問題なし．
(B) 部分訳: 訳文は少し情報が欠けている．
(C) 可能訳: 訳文はかなり情報が欠けている．
(D) 不可訳: 訳文からは，情報が想像もできない．

3.2 翻訳ランクと正解群類似度の関係

本節では，正解文追加類似度計算法による TDMT の評価結果について述べる．図 2 は，類似文検索閾値と類似文検索により得られる目的言語類似文数の平均と標準偏差の関係を示している．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 マッチングを行ない，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 2 より類似文検索閾値の減少にともない，追加される正解文の数が増加していることがわかる．

表 1 は類似文検索閾値が 0.6 の場合に，類似文検索により追加される目的言語類似文の一例である．表中の"/"は，日本語における形態素境界を表している．異なる言い回しであるが，同じような意味の目的言語文が得られていることがわかる．

図 3 に，翻訳ランクと正解群類似度の関係を示す．類似文検索閾値は 0.6 としている．図 3 の横軸は翻訳ランクを表しており，縦軸は正解群類似度を表している．○は各テスト文を表しており，●は各翻訳ランク内での正解群類似度の平均を表している．

図 3 より，主観評価で高いランクである翻訳結果ほど，正解群類似度も大きくなる傾向があることがわかる．

3.3 判別分析によるランク決定

正解群類似度から，翻訳ランクを自動判定するため，翻訳ランクと正解群類似度を用いて判別分析を行なった．

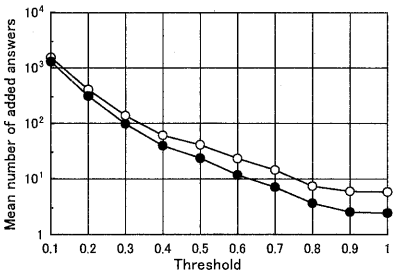

図 2.1 類似文検索閾値と追加される正解文数の平均

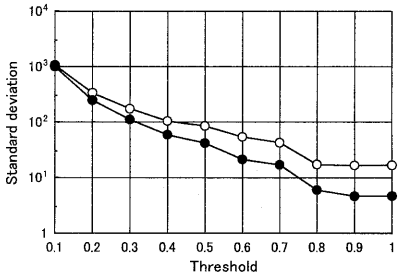

図 2.2 類似文検索閾値と追加される正解文数の標準偏差

図 2 類似文検索閾値と追加される正解文数の関係

表 1 追加される正解文の例

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.

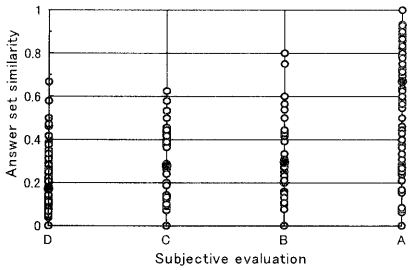

図3　言語翻訳部の翻訳ランクと正解群類似度

3.3.1　判別方法

2 クラス分けの判別と，4 クラス分けの判別を行なった．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，A,B ランクと C,D ランクの 2 クラス分け，A,B,C ランクと D ランクの 2 クラス分けを行なった．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．

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なう．

以下に判別率（Discriminant ratio）を定義する．

$$Dicriminant\ Ratio = \frac{n_{cor}}{n_{total}} \qquad (2)$$

ここで，n_{cor} は正しく判別された文の数であり，n_{total} は，全体の文の数である．

3.3.2　判別結果

図 4 に TDMT の評価結果について判別を行なった結果を示す．図 4 では，類似文検索閾値毎の判別率と，従来の DP マッチングによる類似度を用いて判別を行なった場合の判別率を示している．図 4 において，縦軸は判別率であり，横軸は類似文検索閾値または，従来の DP マッチングによる類似度を用いた方法を表している．また，図中の凡例の "/" は，クラス分けの境界を表している．例えば，A/B,C,D の記述では，A と B,C,D の 2 クラス分けを表している．

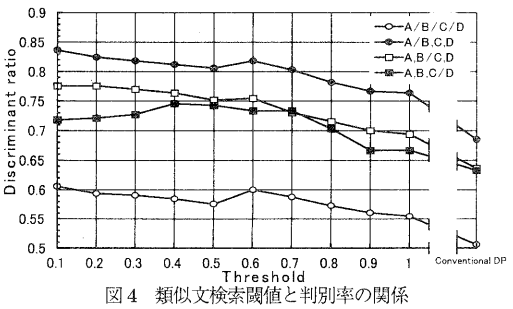

図4　類似文検索閾値と判別率の関係

図 4 では，クラス分けの方法により判別率が最大となる類似文検索閾値は変化するが，全ての類似文検索閾値と全てのクラス分けにおいて，正解群類似度を用いた判別の方が，従来の DP マッチングによる類似度を用いた場合より，判別率が 10%程度改善されている．特に A と B,C,D の 2 クラス分けの判別では効果が大きく，判別率が 68.5%から 83.5%となり 15%の改善がみられている．

3.4　翻訳ランク評価法のコスト削減の検討

前節で述べたように，A と B,C,D の 2 クラス分けにおいては，正解群類似度を用いた判別分析により 83.5%と高い判別率が得られることが分かった．そこで本節では，A ランクの判別を自動で行ない，その他のクラスを主観でランク評価する翻訳評価の効率化について検討する．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，ランク評価結果が A である状態（*class 1*）と，B,C,D ランクである状態(*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(*CLASS2*)の 4 つの組み合わせがあり，表 2 に示す 4 種類の確率が定義できる．

表2　2 クラス判別の 4 つの確率

		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)*

表 2 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(クラス 1 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(クラス 2 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(クラス 1 棄却率：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である．

図 5 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，クラス 2 誤り受理率及びクラス 1 棄却率の関係である．図 5 において横軸は，正解群類似度を表しており，縦軸は，クラス 2 誤り受理率，またはクラス 1 棄却率を表している．これらは，類似文検索閾値は 0.6 とした場合の結果である．図 6 は，従来の DP による類似度を用いた場合について，図 5 と同様にプロットした図である．図 5 及び図 6 において，○はクラス 1 棄却率を表しており，●はクラス 2 誤り受理率を表している．

本方式の性能評価においては，クラス 1 の自動判別が目的であるため，クラス 1 受理率とクラス 2 誤り受理率を評価尺度とする．図 7 は，図 5 及び図 6 から求めた，クラス 1 受理率とクラス 2 誤り受理率との関係である．横軸はクラス 2 誤り受理率，縦軸はクラス 1 受理率である．図 7 の各点は，判別閾値を 0.1 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．○は正解群類似

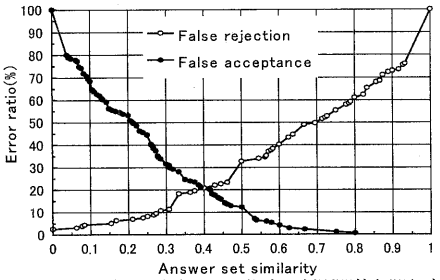

図5　正解群類似度を用いた場合の判別閾値と誤り率の関係

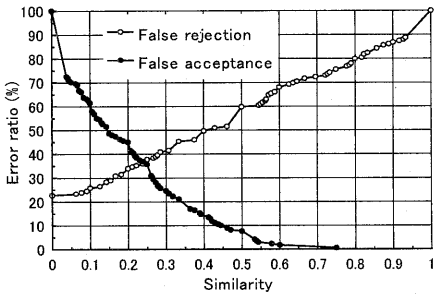

図6　従来のDPによる類似度を用いた場合の判別閾値と誤り率の関係

度を用いた場合の結果であり，●は従来のDPによる類似度を用いた場合の結果である．従来のDPによる類似度を用いた場合と比較し，正解群類似度を用いた場合の方が，全ての点で優れていることが分かる．

次に，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$P(CLASS\ 1)$を求める．

$P(CLASS\ 1)$ は次式で求めることが出来る．

$$P(CLASS\ 1) = P(CLASS\ 1 \mid class\ 1) \times P(class\ 1) + P(CLASS\ 1 \mid class\ 2) \times P(class\ 2) \quad (3)$$

式(3)において，$P(class\ 1)$は全テストセットの内，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内，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0.48，$P(class\ 2)$が0.52である．

式(3)の右辺第一項の$P(CLASS\ 1 \mid class\ 1) \times P(class\ 1)$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$P(CLASS\ 1 \mid class\ 2) \times P(class\ 2)$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．

図8は，$P(CLASS\ 1 \mid class\ 1) \times P(class\ 1)$と，$P(CLASS\ 1 \mid class\ 2) \times P(class\ 2)$の関係を表している．縦軸が$P(CLASS\ 1 \mid class\ 1) \times P(class\ 1)$で，横軸が$P(CLASS\ 1 \mid class\ 2) \times P(class\ 2)$である．図中の○は，正解群類似度を用いた場合の結果であり，●は，従来のDPによる類似度を用いた場合の結果である．

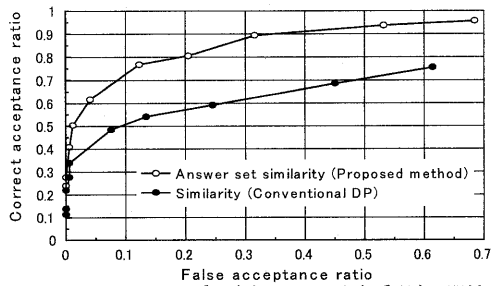

図7　クラス1受理率とクラス2誤り受理率の関係

図8を，テストセット全体に対しての誤り率である横軸を固定してみた場合，正解群類似度を用いた結果の方が，従来のDPによる類似度を用いた場合と比べ，縦軸の値が0.1程度大きくなっている．これは，正解群類似度を用いた場合，従来のDPによる類似度を用いた場合に比べて，全体に対する誤りを同じとしたまま，テストセット全体に対して10%程度多くコストを削減できることを表している．例えば，テストセット全体に対して2.5%の誤りを許容したとすると（図中の破線），提案手法では約30%，従来のDPでは，約20%のコスト削減が可能となっている．

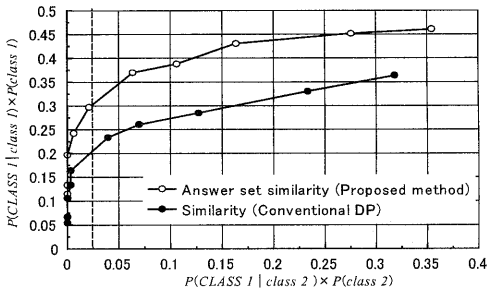

図8　$P(CLASS\ 1 \mid class\ 1) \times P(class\ 1)$と$P(CLASS\ 1 \mid class\ 2) \times P(class\ 2)$の関係

3.5　音声翻訳システムの評価への適用

本節では，正解文追加類似度計算法を，言語翻訳部TDMTの入力側に音声認識部を加えたATR-MATRIX

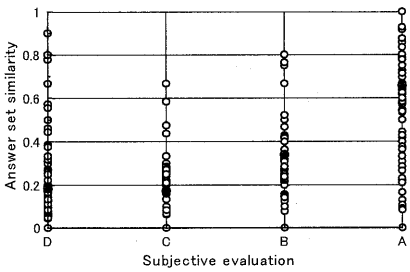

図9　音声認識部を含めた場合の翻訳ランクと正解群類似度の関係

表3　Dランクであるが，正解群類似度が大きくなる例

Example 1	Source language test sentence	コネクティングルームが一泊五万七千円となっております
	Recognition result	コネクティングルームが一泊五〇七千円となっております
	Correct answer	A connecting room is fifty seven thousand yen per night .
	Translation result	A connecting room is five zero seven thousand yen per night .
	Answer set similarity	0.8
Example 2	Source language test sentence	今ワシントンのワシントンホテルに滞在しています
	Recognition result	今ワシントンの足のホテルに滞在しています
	Correct answer	I 'm staying at the washington hotel in Washington .
	Translation result	I 'm staying at the foot hotel in Washington now .
	Answer set similarity	0.78
Example 3	Source language test sentence	二一三五四三の一七五五
	Recognition result	二三五四三の一七五五
	Correct answer	Two one three five four three one seven five five .
	Translation result	Two three five four three , one seven five five .
	Answer set similarity	0.9

の音声翻訳統合評価に適用した結果について示す．

図9に，翻訳ランクと正解群類似度の関係を示す．類似文検索閾値は0.6としている．図9は図3同様，横軸は翻訳ランクを表しており，縦軸は正解群類似度を表している．○は各テスト文を表しており，●は各ランク内での正解群類似度の平均を表している．

図3と図9での結果を比較すると，図9の評価結果では，Dランクと評価されているが，正解群類似度の値が大きいものがある．これらは，表3に示すように音声認識誤りに起因する．すなわち，価格や数字，固有名詞などの情報が誤って音声認識され，それが翻訳されるので，正解群類似度が高いにもかかわらず，重要な情報伝達されていないことから，Dランク判定されているためである．

4. むすび

翻訳システムの新たな評価手法として，正解文追加類似度計算法を提案し，ATR音声翻訳通信研究所で開発されたATR-MATRIXの評価に適用した．この結果を用いて判別分析を行なったところ，従来のDPによる評価結果を用いた場合と比較して，10%程度判別率の改善がみられた．特に，AランクとB,C,Dランクの2クラス分けでは，判別率が68.5%から83.5%となり，15%の改善がみられた．しかし，図3からもわかるように，提案手法を用いても，翻訳ランク評価法では，Aランクと評価されているにもかかわらず，正解群類似度の値が小さいままのものがある．パラフレーズコーパスの利用が可能となれば，より効率的に正解を追加することができ，このようなAランクで正解群類似度が小さい翻訳結果の正解群類似度を大きくすることが出来ると考えられるため，本手法が更に有効になると考えられる．

翻訳ランク評価のコスト削減については，正解群類似度を用いた場合，従来のDPによる類似度を用いた場合と比較して，テストセット全体に対して10%程度多くコストを削減できることが分かった．

音声認識誤りのある場合はDランクであるにもかかわらず，正解群類似度が大きい場合があることが分かった．原因は，ランク評価の評価単位とDPマッチングのマッチング単位のズレである．DPマッチングのマッチング要素は，分かち書きされた単語である．そのため，数字の連鎖からなる電話番号，料金などでは，音声認識により数字が誤っても，翻訳の構造を損なうことなく言語翻訳がなされ，その類似度も高い．しかしながら，翻訳評価では，翻訳結果の数字などの重要な情報が誤っていることからDランクと判定されてしまう．この問題を解決するために，自動翻訳評価に適したマッチング単位を選択する必要があり，今後検討したい．

謝辞　本研究を行なう上で，貴重なご指導を頂いたATR音声言語通信研究所第二研究室　匂坂芳典室長，実験を行なう上で支援して頂いた林輝昭氏に感謝致します．

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．

文献

[1] T. Takezawa, T. Morimoto, Y. Sagisaka, N. Campbell, H. Iida, F. Sugaya, A. Yokoo, S. Yamamoto, "A Japanese-to-English speech translation system: ATR-MATRIX", Proc ICSLP 1998, pp.2779-2782.

[2] E. Sumita, S. Yamada, K. Yamamoto, M. Paul, H. Kashioka, K. Ishilawa, S. Shirai, "Solutions to Problems Inherent in Spoken language Translation : The ATR-MATRIX Approach", Proc. MT Summit'99, Sep.1999.

[3] F. Sugaya,T. Takezawa, A. Yokoo, Y. Sagisaka, S. Yamamoto, "Evaluation of the ATR-MATRIX Speech Translation System with Pair Comparison Method Between the System and Humans", Proc, ICSLP 2000, pp1105-1108.

[4] K. -Y. Su, M. -W. Wu, and J. -S. Chang, "A new quantitative quality measure for machine translation systems", Proc. COLING, pp. 443-439, 1992.

[5] T. Takezawa, F. Sugaya, A. Yokoo, S. Yamamoto, "A New Evaluation Method for Speech Translation Systems and a Case Study on ATR-MATRIX from Japanese to English", Proc. MT Summit'99, Sep.1999.