社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－NL－160　(9)
2004／3／4

日本語話者の英語運用能力との比較による英日翻訳の評価

松吉 俊 †, 安田 圭志 ‡, 竹澤 寿幸 ‡, 菊井 玄一郎 ‡

†京都大学大学院 情報学研究科 知能情報学専攻
‡ATR 音声言語コミュニケーション研究所

翻訳一対比較法の自動化手法は、翻訳自動評価法 (DP ベース自動評価法、 BLEU) を用いて、システムの翻訳能力を Test of English for International Communication (TOEIC) のスコアに換算することができる日英翻訳評価法である。この評価法の利点は、あまりコストをかけずに翻訳一対比較法と非常に近い評価結果を得ることができることである。この翻訳一対比較法の自動化手法を英日方向へ応用する実験を行った。本稿では、翻訳一対比較法の自動化手法を英日方向へ応用する方法を述べた後、提案手法による、英日翻訳システムの評価実験の結果を報告する。

Automatic Evaluation of English-to-Japanese Translation Systems by Measuring Machine Output against Native Japanese Speaker Results

Suguru Matsuyoshi†, Keiji Yasuda‡,
Toshiyuki Takezawa‡, Genichiro Kikui‡

†Department of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
‡ATR Spoken Language Translation Research Laboratories

An automatic scheme for the translation paired comparison method can precisely evaluate a Japanese-to-English speech translation system's capability. This method gives an objective evaluation result, namely a score for the Test of English for International Communication (TOEIC). Also, this method shows a good correlation with evaluation results of the translation paired comparison method. We applied this automatic scheme to evaluate the English-to-Japanese speech translation system. In this paper, we illustrate how this evaluation is done, and describe an automatic evaluation experiment on an English-to-Japanese speech translation system using the proposed method.

1. はじめに

大規模対訳コーパスの開発やコンピュータの高性能化など、周辺技術の発展に伴い、音声対話翻訳技術は、外国人旅行者とホテルのフロント係の会話のような限定された場面での有効性が実証されている[1)]。今後はより広い範囲の会話が扱えるように拡張を可能にする技術の開発を進める必要がある。そのためにはアルゴリズムの研究と並んで、高品質で実際的なコーパスの構築[2)]が重要である。さらに、研究開発を加速させるためには、実際的なコーパスに基づく性能評価技術が不可欠である。

音声対話翻訳技術の総合的な評価手法としては、対話実験による評価が一般的である。対話実験においては、実システムを介した対話が行われるため、システムの処理遅延やGUIに対する利用者の影響など、翻訳能力以外のシステムの性能を測ることができる。この意味において、対話実験による評価は重要であり、システムの開発において必要不可欠なものである。その一方で、課題の設定や話者の個性等、多くの要因が関係する総合的な評価であるため、音声対話翻訳技術の現状を明瞭に把握し、技術の改良にフィードバックする尺度になりにくいという問題があった。

このような背景から、システム性能の数量化を実現する

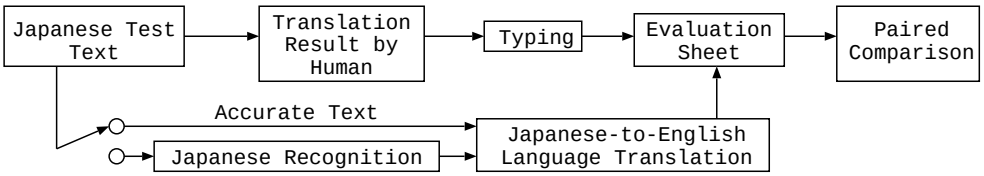

図 1　翻訳一対比較法の処理の流れ
Fig. 1　Diagram of the Translation Paired Comparison Method

手法として、翻訳一対比較法が提案された[3]。翻訳一対比較法は、音声翻訳システムの翻訳能力を Test of English for International Communication (TOEIC)[4] のスコアに換算して表す。この手法は、 TOEIC スコアという、ユーザに対しても明瞭な評価結果を示すことができるという利点を持っている。しかしながら、この翻訳一対比較法は、複数の人間による翻訳結果と、システムによる翻訳結果とを、人手で一対比較する必要があるため、評価に要するコストが大きいという問題があった。

翻訳一対比較法のコストの問題を解決する手法として、翻訳一対比較法の自動化手法が提案されている[5]。翻訳一対比較法の自動化手法は、翻訳自動評価法（DP ベース自動評価法、 BLEU）を用いて、音声翻訳システムの翻訳能力を TOEIC のスコアに換算する。この自動化手法による評価は、翻訳一対比較法による評価と非常に高い相関があることが知られている。この翻訳一対比較法の自動化手法を用いることにより、あまりコストをかけずに明瞭な評価結果を得ることができるようになった。

従来の研究では、まず第一段階として、日英方向について検討と評価実験を行ってきた。しかしながら、これらの翻訳評価法は日英方向に特化した手法ではないため、他の方向の翻訳評価法に応用することができる。また、日本語話者と英語話者の音声対話翻訳を実現するためには、日英方向と英日方向が必須である。そこで、翻訳一対比較法の自動化手法を英日方向の翻訳評価に応用する実験を行った。ただし、単純に日英方向の技術を英日方向に適用すれば良いわけではない。比較対象のデータは日本語ネイティブの TOEIC 被験者によるものである。従来の日英方向の場合は、入力側の日本語音声は完全に聞き取り可能であるのに対し、今回の英日方向の場合は、入力側の英語音声の聞き取り能力にも差がある。そのため、従来の日英方向の場合は、日本語音声を聞かせて英訳させただけであったが、今回は、英語音声の聞き取り、英語音声からの和訳、英語正解テキストからの和訳の 3 種類のデータを集めて、評価分析を実施した。本稿では、翻訳一対比較法の自動化手法を英日方向へ応用する具体的な方法を説明し、提案手法による、英日翻訳システムの評価実験の結果を報告する。

本稿は、以下のように構成される。まず、 2. で翻訳一対比較法とその自動化手法について説明する。次に、 3. において、翻訳一対比較法の自動化手法を英日方向へ応用する方法について述べる。 4. で、その応用手法による、英日翻訳システムの評価実験の結果を報告する。 5. では、前章の結果の考察と、翻訳自動評価法で用いるリファレンス数の検討を行う。最後に、 6. でまとめを述べる。

2. 翻訳一対比較法とその自動化手法

2.1 翻訳一対比較法

翻訳一対比較法は、音声翻訳システムの評価のために提案された評価手法であり、評価結果として、システムの性能を TOEIC スコアに換算する。

図 1 に翻訳一対比較法の処理の流れを示す。ここでの翻訳方向は日英方向である。 6ヶ月以内に TOEIC を受験した複数の日本語ネイティブの被験者に日本語の問題文を音声で提示し、日本語を英語に翻訳させ、解答用紙に記入させる。被験者から回収された解答用紙は書き起こされ、人間の翻訳結果として、システムによる翻訳結果と比較する。これらの比較対象の翻訳結果から評価シートを作成し、日英バイリンガル評価者がシステムと各被験者の翻訳結果とを一対比較で主観評価する。まず、評価者は、システムによる翻訳結果と被験者による翻訳結果に対し、以下のランクを決定する。

(A)　訳文だけで全く問題なし
(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、情報が想像できない

2 つの翻訳の優劣は、ここで決定されたランクに基づき決定するが、 2 つの翻訳のランクが同じであった場合、各翻訳の自然性を考慮して優劣を決定する。自然性まで考慮しても優劣を決めることができない場合については、同等 (Even) と評価する。被験者による翻訳結果には、スペルミスが含まれる場合があるが、本手法の目的は、英語の語彙能力を測定することではなく、音声翻訳能力を測定することであることから、評価者は、スペルミスが含まれる場合でも、音にして意図した単語が推定できる場合については、誤りとしない方針で評価を行っている。

全ての被験者とシステムとの一対比較が完了した段階で、回帰分析によりシステム性能に相当する TOEIC スコア (システム TOEIC 換算点) を計算する。回帰分析には、各被験者の TOEIC スコアと、次式に定義する各被験者ごとの被験者勝率 (W_H) を用いる。

$$W_H = (N_{human} + 0.5 \times N_{even})/N_{total} \qquad (1)$$

ここで、 N_{total} はテストセットに含まれる発話数を、N_{human} は各被験者による翻訳結果がシステムの翻訳結果よりも優れていると評価された発話数を、 N_{even} は各被験

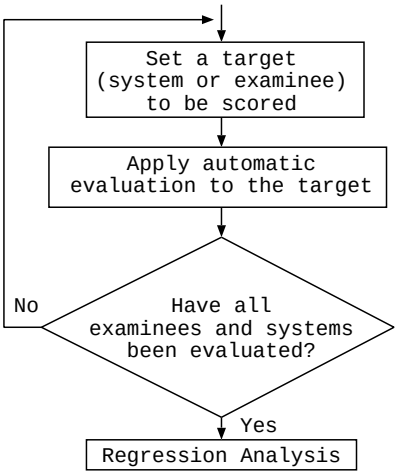

図 2 翻訳一対比較法の自動化手法の評価手順
Fig. 2 The Automatic Schemes for the Translation Paired Comparison Method

者とシステムによる翻訳結果の品質が同等であると評価された発話数を表す。

すべての被験者に対して被験者勝率 (W_H) を計算した後、横軸を TOEIC スコア、縦軸を被験者勝率 (W_H) として値をプロットし、回帰分析を用いて、システム能力と被験者の能力とが均衡する点、即ち $W_H = 0.5$ となる TOEIC スコアを求める。得られた値が、システム TOEIC 換算点である。

2.2 翻訳一対比較法の自動化手法

翻訳一対比較法の自動化手法は、翻訳自動評価法（DP ベース自動評価法、BLEU）を用いて、音声翻訳システムの性能を TOEIC のスコアに換算する。

この手法の処理の流れは、図 1 とほぼ同じであるが、一対比較の代わりに翻訳自動評価法を用いて翻訳結果の評価を行うところが異なる。翻訳一対比較法の自動化手法の評価手順を**図 2** に示す。先程と同じ条件で人間の翻訳結果を収集した後、翻訳自動評価法を用いて被験者の翻訳結果とシステムの翻訳結果にスコアをつけ、回帰分析を行い、システム TOEIC 換算点を求める。

ここでは、翻訳自動評価法として、DP ベース自動評価法と BLEU を用いる。これらの自動評価法は、あらかじめ人手で作成された複数の正解翻訳（リファレンス）と評価対象の翻訳結果を比較することにより、評価対象の翻訳結果にスコアをつける。

以下、DP ベース自動評価法と BLEU について概説する。

2.2.1 DP ベース自動評価法

DP ベース自動評価法は、リファレンスと評価対象の翻訳結果との比較方法として DP マッチングを用いる☆。具体的には、以下に定義するスコア (S_{DP}) により翻訳の評価を行う。

$$S_{DP} = \frac{1}{N_{total}} \sum_{j=1}^{N_{total}} \max_i \left(\frac{T_{ij} - S_{ij} - I_{ij} - D_{ij}}{T_{ij}}, 0 \right) \tag{2}$$

ただし、N_{total} はテストセットに含まれる発話数、T_{ij} はテストセットの発話 j に対するリファレンス i の総語数、S_{ij} はテストセットの発話 j に対するリファレンス i と評価対象の翻訳を DP マッチングにより比較した時の置換語数、I_{ij} は同様に比較した時の挿入語数、D_{ij} は同様に比較した場合の脱落語数である。ここで、置換は、「脱落＋挿入」とは考えない。

2.2.2 BLEU

BLEU[6] は、リファレンスと評価対象の翻訳結果との比較方法として n-gram のマッチ率を計算する。具体的には、以下に定義するスコア (S_{BLEU}) により翻訳の評価を行う。

$$S_{BLEU} = \exp\left\{ \sum_{n=1}^{N} w_n \log(p_n) - \max\left(\frac{L^*_{ref}}{L_{sys}} - 1, 0 \right) \right\} \tag{3}$$

ただし、N は n-gram マッチ率を計算する際に扱う最大 n-gram 長を表し、w_n はその逆数を表す。また、L_{sys} は評価対象の翻訳に含まれる単語数を表し、L^*_{ref} は、L_{sys} と最も単語数が近いリファレンスに含まれる単語数を表す。p_n は修正 n-gram 適合率であり、式 (4) により計算される。

$$p_n = \frac{\sum_i \left(\begin{array}{c} \text{the number of } n\text{-grams in} \\ \text{segment } i \text{ in the translation being} \\ \text{evaluated, with a matching reference} \\ \text{co-occurrence in segment } i \end{array} \right)}{\sum_i \left(\begin{array}{c} \text{the number of } n\text{-grams in segment } i \\ \text{in the translation being evaluated} \end{array} \right)} \tag{4}$$

本研究では、すべて最大 n-gram 長を 2 として BLEU を用いている。

3. 英日翻訳評価法への応用

翻訳一対比較法の自動化手法を英日翻訳システムの評価に応用する。

図 3 に英日方向への応用手法の処理の流れを示す。6ヶ月以内に TOEIC を受験した複数の日本語ネイティブの被験者に英語文を音声またはテキストで提示し、それを日本語に翻訳させ、端末から入力させる。被験者の翻訳結果とシステムの翻訳結果に翻訳自動評価法を用いてスコアをつけ、回帰分析を行い、システム TOEIC 換算点を求める。

回帰分析の信頼性を保証するため、翻訳自動評価法によ

☆ 本研究では、日本語、英語ともに ATR の形態素解析ツールを用いた。

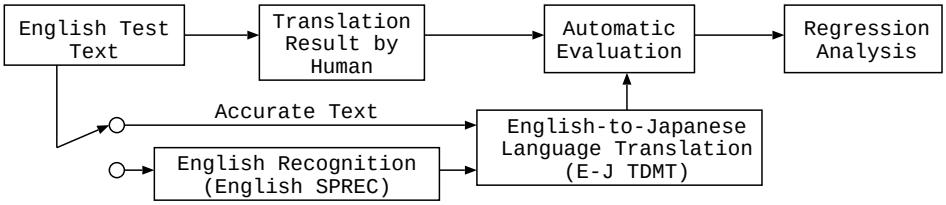

図 3 応用手法の処理の流れ
Fig. 3 Diagram of the Automatic Evaluation Method Applied to an English-to-Japanese Speech Translation System

表 1 翻訳自動評価法によるスコアと TOEIC スコアとの相関
Table 1 Correlations between Automatic Measures and TOEIC scores

リファレンス数	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
音声 +DP	.762	.800	.808	.806	.792	.801	.788	.796	.780	.793	.792	.790	.790	.793	.791
テキスト +DP	.677	.757	.748	.760	.737	.756	.736	.753	.732	.744	.747	.741	.744	.744	.745
音声 +BLEU	.714	.726	.738	.744	.746	.741	.748	.752	.746	.750	.746	.750	.750	.751	.752
テキスト +BLEU	.495	.545	.577	.602	.608	.610	.637	.635	.627	.641	.638	.647	.649	.658	.659

る被験者の翻訳結果のスコアと被験者の TOEIC スコアが相関を持っていることが、この手法を用いる上での不可欠な前提条件となる。この前提条件が満たされていない場合、応用手法を用いて算出された評価結果は信用できない。それ故、テストセットとリファレンスを注意深く設定する必要がある。

本研究では、機械翻訳介在バイリンガル音声対話コーパス MAD1[2] の英語ネイティブの発話から 504 発話を選択し、それをテストセットとした。テストセットにおける 1 発話あたりの平均単語数は 9.7、単語数の標準偏差は 6.1 である。また、1 発話あたりの最小単語数は 1、最大単語数は 33 である。

翻訳自動評価で必要となるリファレンスは、1 発話につき 15 のリファレンスを人手で作成している。

本研究における、翻訳を作成する被験者は、TOEIC スコア 300 点台から 900 点台の計 20 名であり、900 点台を除いて、100 点台ごとに 3 名となっている。

応用手法による評価の信頼性を保証するため、これらの被験者に対して上で述べた相関の度合い (相関係数) を測定した。被験者の翻訳結果には入力ミスなどが含まれている。翻訳自動評価法による評価の際、スコアが不当に低くなる原因になるので、それらを人手で修正している。

リファレンス数を変化させたときの相関係数を**表 1** に示す。表の「音声」、「テキスト」は、それぞれ人間の翻訳結果として英語音声の和訳、英語正解テキストの和訳を用いたことを示す。「+DP」、「+BLEU」は、それぞれ翻訳自動評価法として DP ベース自動評価法、BLEU を用いて翻訳結果のスコアを算出したことを示す。表 1 をグラフにしたものが**図 4** である。棒グラフは左から順に「音声 +DP」、「テキスト +DP」、「音声 +BLEU」、「テキスト +BLEU」を表している。

これらを見ると、「音声 +DP」、「テキスト +DP」、「音声 +BLEU」においては高い相関があることが分かる。また、「テキスト +BLEU」においても、リファレンス数を大きく取れば、相関があることが分かる。従って、応用手法による評価結果は十分信頼するに値すると言える。

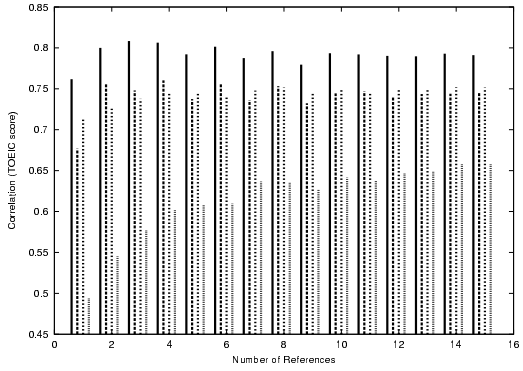

図 4 翻訳自動評価法によるスコアと TOEIC スコアとの相関
Fig. 4 Correlations between Automatic Measures and TOEIC scores
左から順に「音声 +DP」、「テキスト +DP」、「音声 +BLEU」、「テキスト +BLEU」。

4. 英日音声翻訳システムの評価実験

前章で述べた応用手法を用いて英日音声翻訳システムの評価実験を行った。翻訳自動評価に用いるリファレンスは 1 発話あたり 15 とした。

本実験で評価対象とするシステムは、音声翻訳システム ATR-MATRIX[1] の言語翻訳サブシステム Transfer Driven Machine Translation (TDMT)[7] を拡張し、機械翻訳介在バイリンガル音声対話の収集に利用したバージョンである。システムの入力は、英語音声をタイピストがタイプしたものとした。タイプミスは極めて少ないため、システムにはテストセットのテキストに近いものが入力されたと考えてよい。

システムに英語のテキストを翻訳させ、翻訳自動評価法を用いてその翻訳結果にスコアをつけた。DP ベース自動評価法によるスコアは 0.346 であった。BLEU によるスコ

表 2　リファレンス数に対するシステム TOEIC 換算点
Table 2　System's TOEIC scores to Number of References

リファレンス数	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
音声　+DP	339	401	389	420	414	413	398	416	423	416	411	419	411	418	419
テキスト +DP	118	227	196	252	231	236	205	235	240	226	222	229	221	229	233
音声　+BLEU	754	731	724	721	716	719	711	705	709	703	705	702	709	701	704
テキスト +BLEU	589	532	509	512	487	498	477	456	467	451	450	441	461	441	448

アは 0.809 であった。回帰分析により、これらのスコアからシステム TOEIC 換算点を求める。

被験者の翻訳結果のスコアに対して回帰直線を引いたものを**図 5**、**図 6** に示す。図 5、図 6 では、翻訳自動評価法として DP ベース自動評価法、BLEU がそれぞれ用いられている。これらの図における水平の直線は、システムの翻訳結果のスコアを表す。

異なる条件の下で算出された、それぞれのシステム TOEIC 換算点を**表 2** のリファレンス数 15 の欄に示す。人間の翻訳結果として英語正解テキストの和訳を用いている「テキスト」の欄には、音声翻訳システムの言語翻訳サブシステムの TOEIC 換算点が示されている。一方、この実験においては音声認識器を用いて英語音声を認識することをせず、テキストをシステムの入力としているため、人間の翻訳結果として英語音声の和訳を用いている「音声」の欄には、音声認識器の認識精度が 100% に近い状態における音声翻訳システムの TOEIC 換算点が示されている。日本語ネイティブの英語聞き取り能力は 100% 完璧ではないので、英語音声の和訳の翻訳品質は、英語正解テキストの和訳の翻訳品質よりも低くなる。このため、表 2 における「音声」の欄の値は、同じ翻訳自動評価法を用いている「テキスト」の欄の値よりも大きい値となっている。

参考に、日本語ネイティブの英語聞き取り能力を DP ベース自動評価法と BLEU で求めた。結果を**図 7** に示す。DP ベース自動評価法の類似度と TOEIC スコアの相関係数は 0.850、BLEU による類似度と TOEIC スコアの相関係数は 0.853 であった。図 7 とこれらの値から、TOEIC スコアが高いほど、テストセットの聞き取りの成績が良いことが分かる。さらに、回帰直線を求めれば、英語音声認識システムの性能を TOEIC スコアに換算できる可能性もあるが、本稿ではこれ以上の詳しい議論は行わない。

5. 考察と検討

5.1 実験結果の考察

前章で得られたシステム TOEIC 換算点は、翻訳自動評価法の選択によって大幅に値が異なる。BLEU を用いて算出したシステム TOEIC 換算点は、DP ベース自動評価法を用いて算出したシステム TOEIC 換算点よりも高い値となっている。この原因について考察する。

BLEU を用いて算出したシステム TOEIC 換算点が高い理由は、BLEU による、被験者の翻訳結果のスコアが相対的に低いためであると考えられる。

システムは、どのようなテキストが入力されたとしても、常にある程度の出力を返すように設計されている。一方で、被験者は、翻訳が困難な部分を含む発話に遭遇した場合、翻訳が可能な部分のみを訳出し、翻訳が困難な部分に対してはあまり訳出しないということが多い。**表 3** に、被験者の英語音声の書き取りの 1 発話あたりの平均単語数、および、英語音声の和訳、英語正解テキストの和訳の 1 発話あたりの平均形態素数を TOEIC スコアとともに示す。なお、システムの翻訳結果の 1 発話あたりの平均形態素数は 12.6 であった。

BLEU は、評価対象の翻訳結果に含まれる単語数が少ない場合、ペナルティーを課す。次式が BLEU スコアのペナルティーを与える式である。(式 (3) 参照)

$$S_{penalty} = \exp\left\{\max\left(\frac{L^*_{ref}}{L_{sys}} - 1, 0\right)\right\} \qquad (5)$$

被験者の英語音声の和訳と英語正解テキストの和訳に対する BLEU スコアのペナルティー ($S_{penalty}$) を**図 8** に示す。システムの翻訳結果に対する BLEU スコアのペナルティー ($S_{penalty}$) は 1.00 であった。

図 8 から、被験者は、翻訳困難な部分をあまり訳出せず、相対的に大きなペナルティーを受けているということが見て取れる。故に、ある程度の出力をしているためペナルティーが少ないシステムは、BLEU において高く評価された結果、高いシステム TOEIC 換算点を与えられていると考えられる。

日英方向においては、翻訳一対比較法による評価結果は、DP ベース自動評価法による評価と BLEU による評価の中間の値となった[5]。一方、英日方向に対して行った小規模な予備実験によると、翻訳一対比較法による評価は、DP ベース自動評価法による評価と同程度の低い値となる傾向がある。日本語ネイティブとシステムの翻訳結果に同程度の情報が含まれている場合、たいてい日本語ネイティブの和訳の方がシステムの翻訳結果より自然性が高い。そのため、翻訳一対比較法によるシステムの評価は、2 つの自動評価法による評価の中間の値とならず、それらよりも低い値になると考えられる。

5.2 リファレンス数の検討

翻訳自動評価法は、その評価においてリファレンスを必要とする。リファレンスの作成にはコストがかかるため、使用するリファレンスの数はなるべく少なくしたい。しかし、あまりにもリファレンス数が少ないと、翻訳自動評価における評価結果が信頼できなくなる恐れがある。

以下、英日方向においてリファレンス数がどれほどあれば信頼できる評価が得られるかについて検討する。

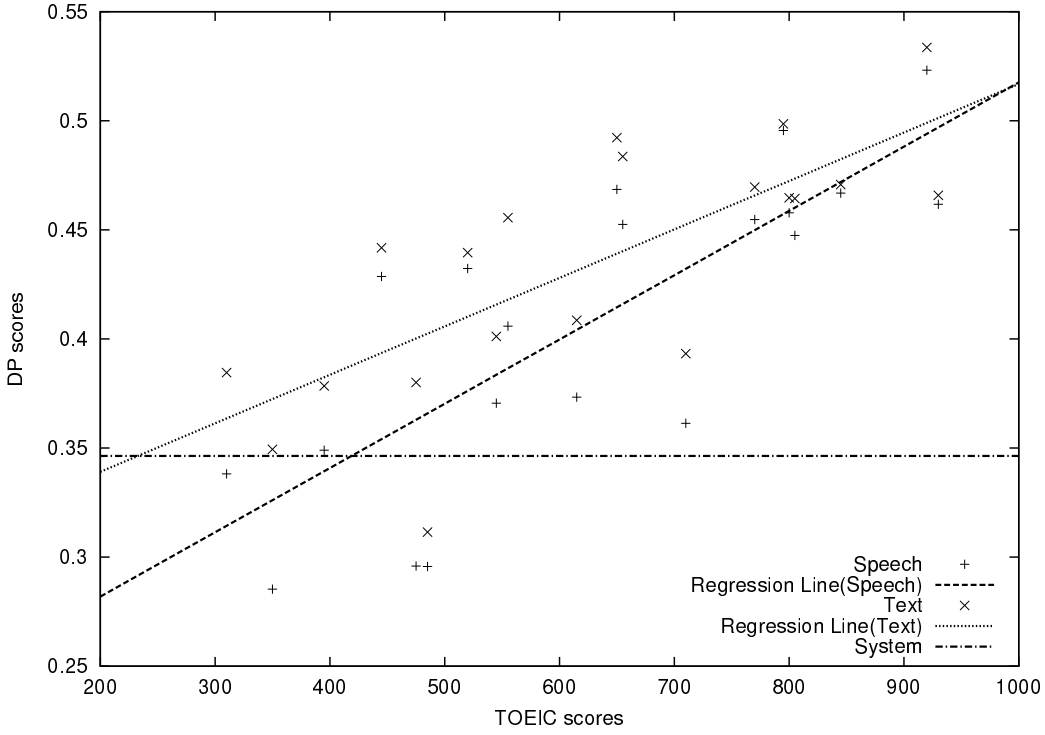

図 5 DP ベース自動評価法による、被験者の翻訳結果のスコアの回帰直線

Fig. 5 Regression Lines of DP scores of the Subjects

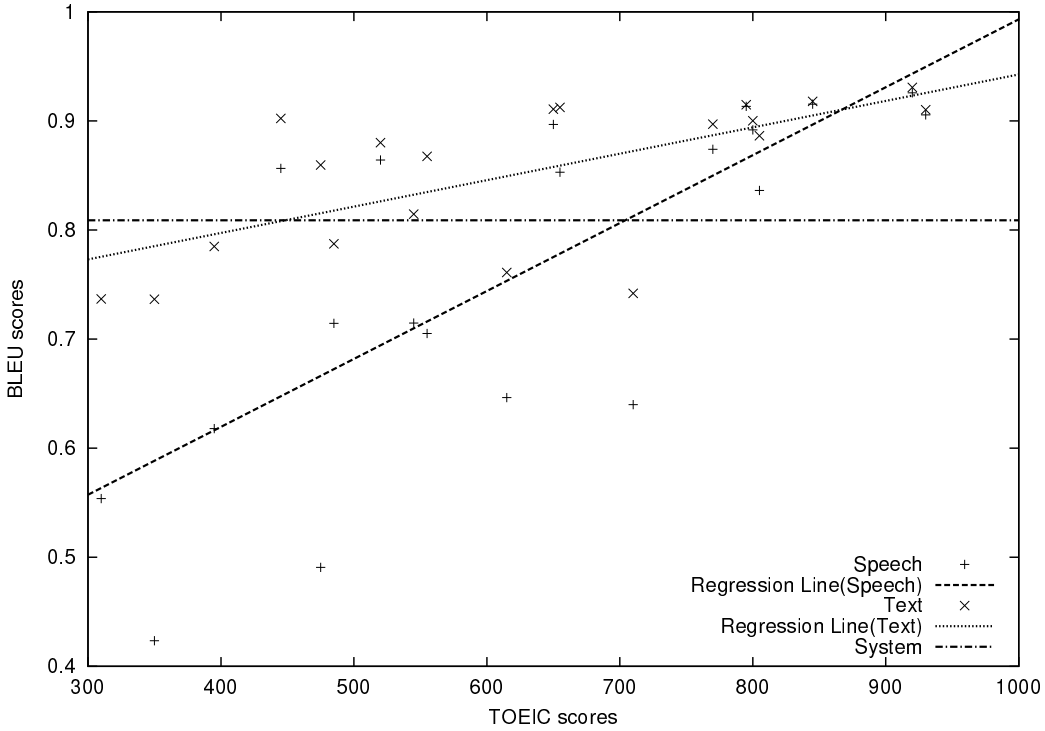

図 6 BLEU による、被験者の翻訳結果のスコアの回帰直線

Fig. 6 Regression Lines of BLEU scores of the Subjects

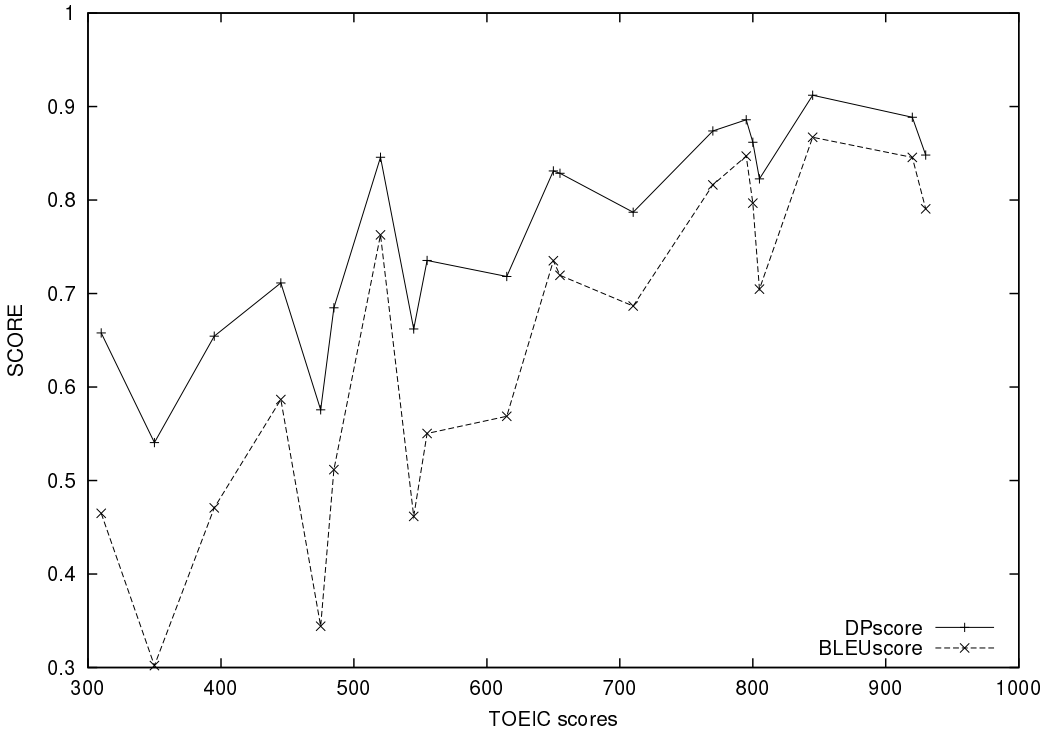

図 7　日本語ネイティブの英語聞き取り能力
Fig. 7　English Dictation Written by Native Japanese Speakers

表 3　各被験者の 1 発話あたりの平均単語数 (英語) / 平均形態素数 (日本語)
Table 3　The Average Number of Words per Utterance when Subjects write the Utterances and the Translations

TOEIC スコア	310	350	395	445	475	485	520	545	555	615
英語音声の書き取り	7.4	5.9	7.4	8.7	5.8	7.5	10.1	6.8	7.6	8.3
英語音声の和訳	7.2	6.2	8.0	11.8	6.9	10.4	11.3	9.2	9.1	8.3
英語正解テキストの和訳	9.3	10.3	10.2	13.4	12.8	12.1	11.6	10.7	11.8	9.7
TOEIC スコア	650	655	710	770	795	800	805	845	920	930
英語音声の書き取り	9.8	9.3	9.0	10.1	10.3	9.9	9.3	10.4	10.5	10.0
英語音声の和訳	12.2	11.5	8.0	11.9	12.9	11.4	11.5	14.3	12.6	12.1
英語正解テキストの和訳	12.7	12.6	9.1	12.4	13.0	11.6	12.6	14.4	12.7	12.2

前章の評価実験と同じ条件の下、リファレンス数のみを変化させて、システム TOEIC 換算点を求めた。結果を表 2 に示す。この表をグラフにしたものが図 9 である。棒グラフは左から順に「音声 +DP」、「テキスト +DP」、「音声 +BLEU」、「テキスト +BLEU」を表している。

図 9 から、リファレンス数が 10 以上になると、リファレンス数が増加しても、システム TOEIC 換算点はほとんど増減しないということが見て取れる。この事実より、信頼できる評価計算には、リファレンス数は 10 程度で十分であるということが結論づけられる。

6. おわりに

翻訳一対比較法の自動化手法を英日方向に応用する実験を行った。評価実験の結果、日英方向で確立された手法を英日方向にも応用できる見込みがあることが分かった。また、英日方向の翻訳自動評価法に必要となるリファレンスは、10 程度あれば十分信頼のできる評価が得られることが分かった。残された課題としては、例えば別のテストセットに対して本手法を用い、日英方向のときのように 2 つの異なる翻訳自動評価法によって得られるシステムの評価が近い値となるかどうかを検証し、本手法の有効性や普遍性を議論する必要がある。

今後は、日英方向、英日方向ともに、音声認識結果を入力とした場合の音声翻訳能力を低コストで推定する手法を研究する予定である。さらに、実際的なコーパス構築とシステム性能の数量化のみならず、翻訳システムの長所や短所の診断が可能な方向を探り、そのためのテストセット選択法や技術について検討する計画である。

謝　辞

本研究は通信・放送機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものであ

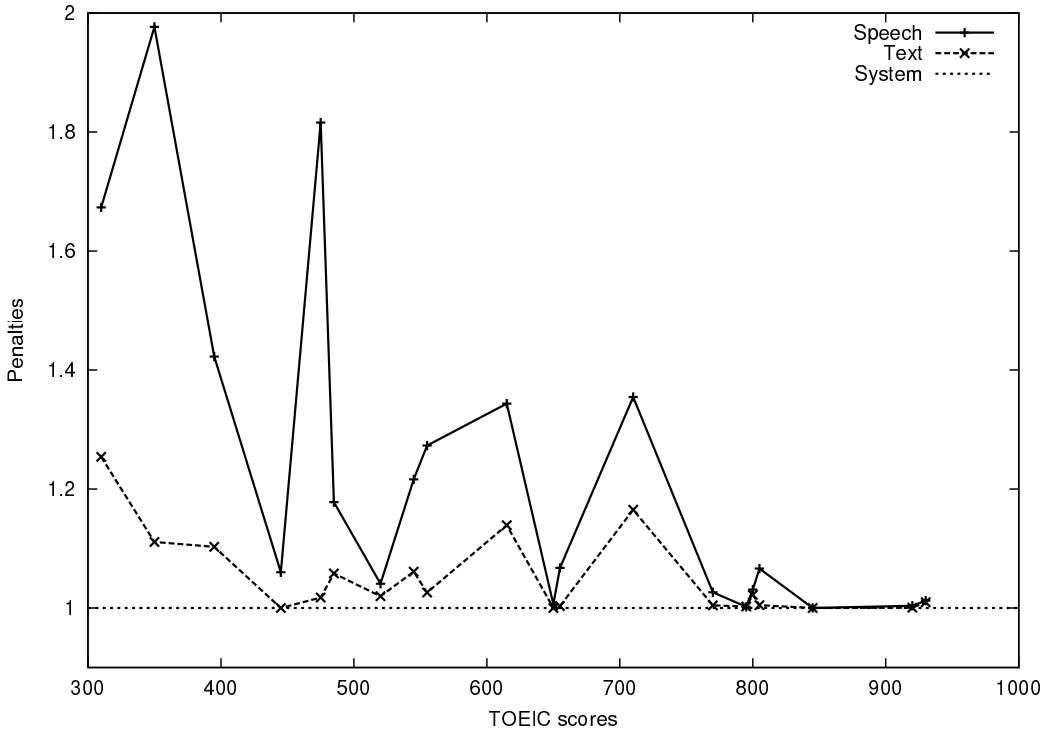

図 8 被験者の BLEU スコアのペナルティー
Fig. 8 Penalties in BLEU scores of the Subjects

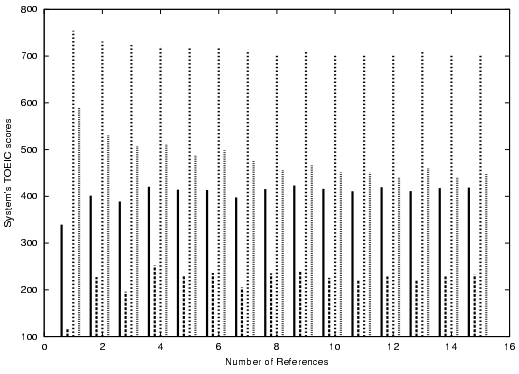

図 9 リファレンス数に対するシステム TOEIC 換算点
Fig. 9 System's TOEIC scores to Number of References
左から順に「音声 +DP」、「テキスト +DP」、「音声 +BLEU」、「テキスト +BLEU」。

る。

参 考 文 献

1) 菅谷史昭, 竹澤寿幸, 隅田英一郎, 匂坂芳典, 山本誠一: 音声翻訳システム:ATR-MATRIX の開発と評価, 情報処理学会論文誌, Vol. 43, No. 7, pp. 2230–2241 (2002).
2) Takezawa, T. and Kikui, G.: Collecting machine-translation-aided bilingual dialogues for corpus-based speech translation, *Proceedings of 8th European Conference on Speech Communication and Technology(EUROSPEECH)*, Vol. 4, pp. 2757–2760 (2003).
3) 菅谷史昭, 竹澤寿幸, 横尾昭男, 山本誠一: 音声翻訳システムと人間との比較による音声翻訳能力評価手法の提案と比較実験, 電子情報通信学会論文誌, Vol. J84-D-II, No. 11, pp. 2362–2370 (2001).
4) TOEIC: Test of English for International Communication(2002), http://www.toeic.com/.
5) Yasuda, K., Sugaya, F., Takezawa, T., Yamamoto, S. and Yanagida, M.: Applications of Automatic Evaluation Methods to Measuring a Capability of Speech Translation System, *Proceedings of 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 371–378 (2003).
6) Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: Bleu: a Method for Automatic Evaluation of Machine Translation, *Proceedings of the Association for Computational Linguistics(ACL)*, pp.311–318 (2002).
7) 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).