Internal Use Only（非公開）

TR－SLT－0051

英日方向の音声翻訳能力評価実験

An Experiment for Automatic Evaluation of the English-to-Japanese Speech Translation System

松吉　俊　　竹澤寿幸
Suguru Matsuyoshi　　Toshiyuki Takezawa

２００３年９月１９日

概要

日英方向の翻訳評価法が英日方向にも応用できる見込みがあるということを示す。現在、日英方向の翻訳評価法としては様々な手法が提案されているが、英日方向の定量的でわかりやすい翻訳評価法は存在しなかった。そこで、日英方向の翻訳評価法である、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。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)のスコアとして換算することができる翻訳評価法であり、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という優れた長所を持つ。本稿では、はじめに、TOEIC_MAD_EJ データについて説明し、その解析結果を報告する。次に、TOEIC_MAD_EJ データとリファレンス（複数の正解翻訳）を用いて翻訳一対比較法の自動化手法を英日方向に応用する方法について述べ、応用手法による、TDMT の MAD1 version の評価実験の結果を報告する。

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目２番地２ TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai “Keihanna Science City” 619-0288,Japan
Telephone:+81-774-95-1301
Fax　　:+81-774-95-1308

目次

1 はじめに

音声翻訳システムの研究・開発において、翻訳評価法は必要不可欠なものである。低コストで客観的な評価結果を与えるような翻訳評価法は優れた翻訳評価法であり、そのような翻訳評価法は大いに望まれている。

日英方向の音声翻訳システムの翻訳評価法に、翻訳一対比較法がある [1]。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) [2] のスコアに換算して表すことができる。この手法は、TOEIC スコアという明瞭な評価結果を示すことができるが、その評価に非常にコストがかかる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た [3]。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。この手法によって、あまりコストをかけずに、翻訳一対比較法の評価結果と非常に近い評価結果を得ることができるようになった。

しかし、上記の翻訳評価法は、日英方向の音声翻訳システムの評価に限られていた。翻訳一対比較法の自動化手法は日英方向の優れた翻訳評価法である。この手法を英日方向の翻訳評価に応用できないだろうか。

本論文では、日英方向の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可能性について検討した結果を報告する。

本論文は、以下のように構成される。まず、2. で翻訳一対比較法の自動化手法について、その概要を述べる。次に、3. で、テストセットと被験者の翻訳結果の集合である TOEIC_MAD_EJ データについて説明し、その解析結果を示す。4. では、その TOEIC_MAD_EJ データとリファレンスを合わせて解析した結果を示す。5. で、実際に翻訳一対比較法の自動化手法を英日翻訳システムの翻訳評価に応用した実験について述べ、その結果を示す。6. では、前章の結果の考察、および新たな翻訳自動評価法やリファレンス数の考察を行う。7. はまとめである。

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について概要を述べる。

翻訳一対比較法の自動化手法は、これまでに提案されている翻訳自動評価法を用いて TOEIC スコアが異なる複数の被験者の翻訳結果と翻訳システムの翻訳結果にスコアをつけ、回帰分析によりシステム TOEIC 換算点を求める評価手法である。この過程を図 1に示す。

翻訳一対比較法の自動化手法では、翻訳自動評価法として DP ベース自動評価法と BLEU を使用している。以下、これら 2 つの翻訳自動評価法について概説する。

2.1 DP ベース自動評価法

DP ベース自動評価法では、以下に定義するスコア S_{DP} により翻訳の評価を行う。

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

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 BLEU

BLEU では、以下に定義するスコア S_{BLEU} により翻訳の評価を行う。

$$S_{BLEU} = \exp\left\{ \sum_{n=1}^{N} w_n \log(p_n) - \max\left(\frac{L^*_{ref}}{L_{sys}} - 1, 0 \right) \right\} \tag{2}$$

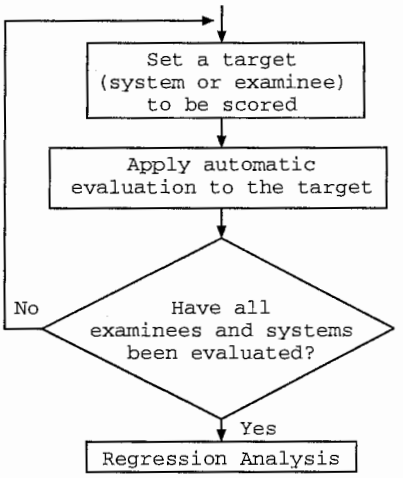

図 1: 翻訳一対比較法の自動化手法の評価手順

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、M_i^{n-gram} をテストセットの発話 i を翻訳した文の n-gram のうちリファレンスとマッチするものの数、G_i^{n-gram} をテストセットの発話 i を翻訳した文の n-gram の数としたとき、次式により計算される。

$$p_n = \frac{\sum_i M_i^{n-gram}}{\sum_i G_i^{n-gram}} \tag{3}$$

3 TOEIC_MAD_EJ データ

TOEIC_MAD_EJ データは、テストセットの英語の発話と被験者のテストセットの和訳文を集めたものである。以下、このデータについて詳細に説明し、その後、このデータから得られる解析結果について述べる。

3.1 TOEIC_MAD_EJ データとは

TOEIC_MAD_EJ データの作成過程について詳説する。TOEIC_MAD_EJ データは次のようにして作成された。

- MAD1 における英語の native speaker の発話から 504 の発話を選択した
  (以下、その 504 の発話をテストセットと呼ぶ)
- TOEIC スコアが異なる被験者 20 名を集め、以下の課題をさせた
  1. テストセットの英語の発話を聞いて、それを書き取る
     (以下、被験者が聞取った英語をテキストにしたものを、「聞取った英語の発話」と呼ぶ)
  2. 「聞取った英語の発話」を和訳する
     (以下、その和訳文を「聞取った英語の発話の和訳文」と呼ぶ)、
  3. テキストとして示されたテストセットの発話 (以下、「正解の英語の発話」と呼ぶ) を見て、それを和訳する
     (以下、その和訳文を「正解の英語の発話の和訳文」と呼ぶ)

発話数	504
1発話あたりの単語数の平均	9.7
1発話あたりの単語数の標準偏差	6.1
1発話あたりの最小の単語数	1
1発話あたりの最大の単語数	33

表 1: テストセットのデータ

被験者の TOEIC スコア	310	350	395	445	475	485	520	545	555	615
「聞取った英語の発話」	7.4	5.9	7.4	8.7	5.8	7.5	10.1	6.8	7.6	8.3
「聞取った英語の発話の和訳文」	7.2	6.2	8.0	11.8	6.9	10.4	11.3	9.2	9.1	8.3
「正解の英語の発話の和訳文」	9.3	10.3	10.2	13.4	12.8	12.1	11.6	10.7	11.8	9.7
被験者の TOEIC スコア	650	655	710	770	795	800	805	845	920	930
「聞取った英語の発話」	9.8	9.3	9.0	10.1	10.3	9.9	9.3	10.4	10.5	10.0
「聞取った英語の発話の和訳文」	12.2	11.5	8.0	11.9	12.9	11.4	11.5	14.3	12.6	12.1
「正解の英語の発話の和訳文」	12.7	12.6	9.1	12.4	13.0	11.6	12.6	14.4	12.7	12.2

表 2: 各被験者の、それぞれのデータの一発話あたりの平均単語数

上の「正解の英語の発話」、「聞取った英語の発話」、「聞取った英語の発話の和訳文」、および「正解の英語の発話の和訳文」を発話ごとに集めて整理したものが、TOEIC_MAD_EJ データである。ここでは整理として、音にしては意味が取れるが、綴りが誤っているという単語のスペルミスの修正、および、何か単語が存在することを示すために被験者が独自に使用した特殊記号の削除を人手で行っている。

テストセットのデータを表 1に示す。また、被験者は、TOEIC 試験を受けてから半年以内の、TOEIC スコア 300 点台から 900 点台の人 20 名であり、900 点台を除いて 100 点台ごとに 3 名ずつとなっている。

各被験者の「聞取った英語の発話」、「聞取った英語の発話の和訳文」、「正解の英語の発話の和訳文」の一発話あたりの単語数の平均を表 2に示す。

以下、表および図の中では、「聞取った英語の発話の和訳文」を hJ、「正解の英語の発話の和訳文」を cJ と略記することにする。

3.2 TOEIC_MAD_EJ データの解析

上のようにして作成した TOEIC_MAD_EJ データの「聞取った英語の発話」、「聞取った英語の発話の和訳文」、「正解の英語の発話の和訳文」がどのような特徴を持っているかということを明らかにするために、データの解析を行った。以下、用いた解析手法とその解析結果について述べる。

はじめに、「正解の英語の発話の和訳文」の Perplexity と TOEIC スコアの関係、「聞取った英語の発話の和訳文」の Perplexity と TOEIC スコアの関係について調査した (Perplexity については Appendix を参照のこと)。

word Perplexity を用いて得られた結果を図 2に示す。sentence Perplexity を用いて得られた結果を図 3に示す。これらの図を見比べると、sentence Perplexity の方が TOEIC スコアとの間に正の相関があることが分かる。

次に、以下のようにデータを 2 組に分けて解析を行った。

- 「正解の英語の発話」と「聞取った英語の発話」　(英語の発話の組)
- 「正解の英語の発話の和訳文」と「聞取った英語の発話の和訳文」　(和訳文の組)

それぞれの組について解析した結果を述べた後、その 2 つの組の間の関係について解析した結果を述べる。

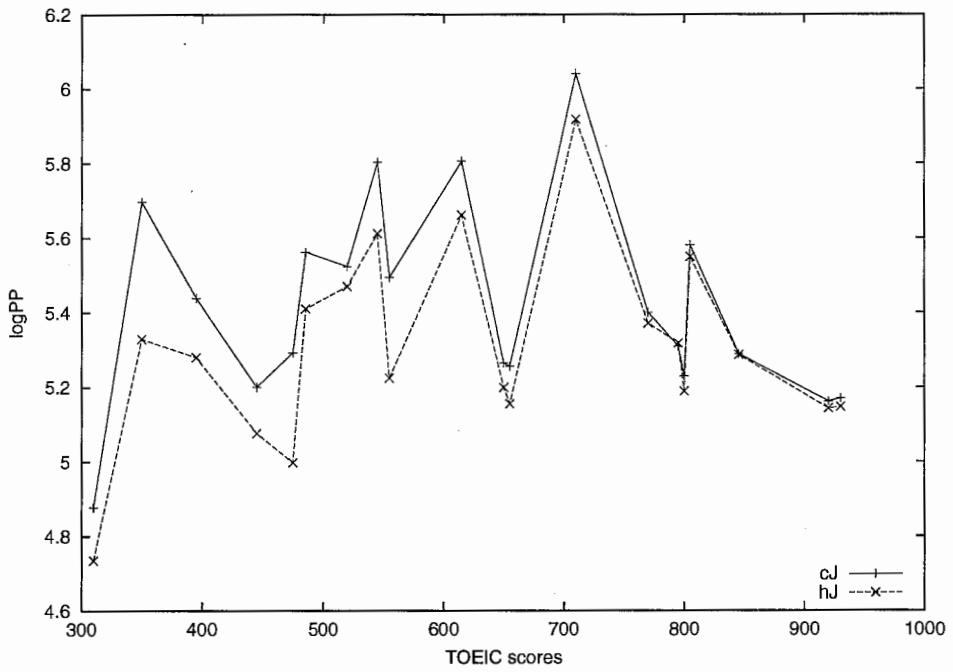

図 2: 「正解の英語の発話の和訳文」、「聞取った英語の発話の和訳文」の word Perplexity

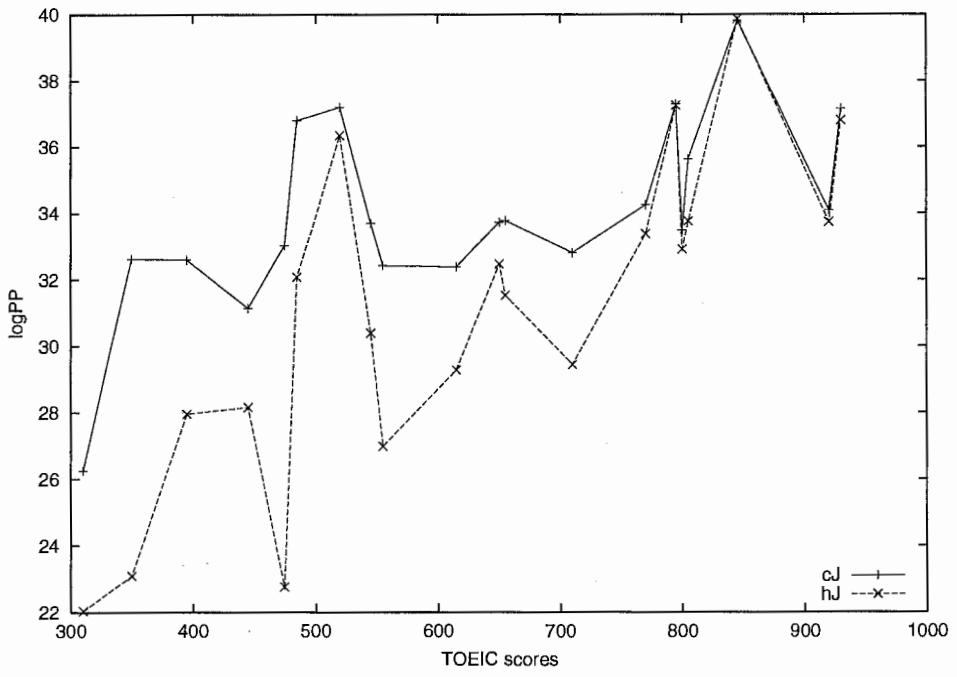

図 3: 「正解の英語の発話の和訳文」、「聞取った英語の発話の和訳文」の sentence Perplexity

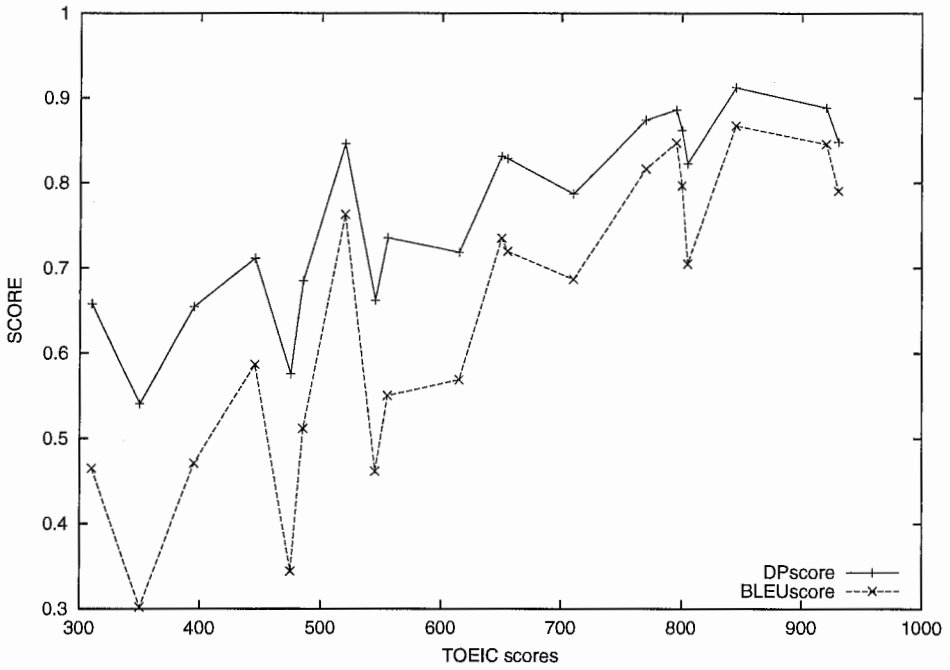

図 4: 「正解の英語の発話」と「聞取った英語の発話」の類似度

3.2.1 「正解の英語の発話」と「聞取った英語の発話」の解析

「正解の英語の発話」と「聞取った英語の発話」の類似度を求め、それと TOEIC スコアとの関係を調査した。類似度の計算には、DP マッチングと最大 n-gram 長を 2 とした BLEU を用いた。結果を図 4に示す。

DP マッチングによる類似度と 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 4とこれらの値より、TOEIC スコアが高いほど、テストセットの発話の聞き取りの成績が良いことが分かる。

3.2.2 「正解の英語の発話の和訳文」と「聞取った英語の発話の和訳文」の解析

「正解の英語の発話の和訳文」と「聞取った英語の発話の和訳文」の類似度を求め、それと TOEIC スコアとの関係を調査した。類似度の計算には、DP マッチング、最大 n-gram 長を 2 とした BLEU、Perplexity の比 (エントロピーの差) を用いた (ここでは、word Perplexity を用いた。Perplexity の比については Appendix を参照のこと)。結果を図 5に示す。

DP マッチングによる類似度と TOEIC スコアの相関係数は 0.746、BLEU による類似度と TOEIC スコアの相関係数は 0.743、Perplexity の比と TOEIC スコアの相関係数は 0.759 であった。図 5とこれらの値より、TOEIC スコアが高いほど、「正解の英語の発話」を見て「聞取った英語の発話の和訳文」を修正する量が少ないことが分かる。

3.2.3 「正解の英語の発話」と「聞取った英語の発話」の類似度と「正解の英語の発話の和訳文」と「聞取った英語の発話の和訳文」の類似度の関係

図 4と図 5を見ると、「正解の英語の発話」と「聞取った英語の発話」の類似度と、「正解の英語の発話の和訳文」と「聞取った英語の発話の和訳文」の類似度は、TOEIC スコアに対して同じような傾向を示している。実際にこの 2 つの類似度がどれくらい相関しているのかを明らかにするために相関係数を算出した。得られた類似度間の相関係数を表 3に示す。表中の (E) は英語の発話の対を対象とした類似度であることを、(J) は和訳文の対を対象とした類似度であることを示す。

この表を見ると、「正解の英語の発話」と「聞取った英語の発話」の類似度と「正解の英語の発話の和訳文」と「聞取った英語の発話の和訳文」の類似度の相関がかなり高いということが分かる。ここから、英語の

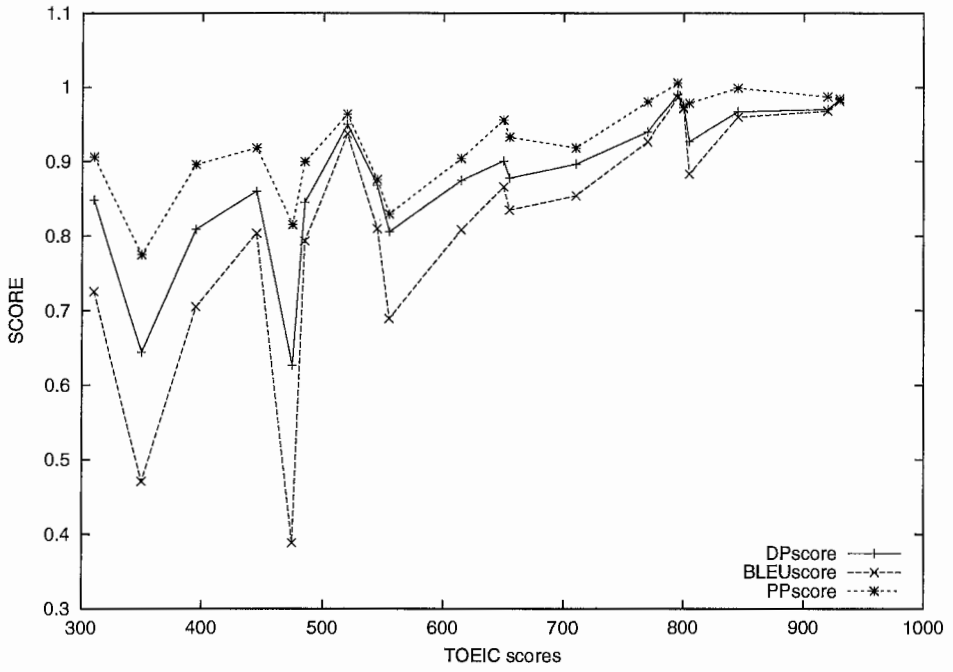

図 5: 「正解の英語の発話の和訳文」と「聞取った英語の発話の和訳文」の類似度

	DPscore(E)	BLEUscore(E)	DPscore(J)	BLEUscore(J)	PPscore(J)
DPscore(E)	1	0.995	0.908	0.906	0.916
BLEUscore(E)	0.995	1	0.911	0.912	0.929
DPscore(J)	0.908	0.911	1	0.994	0.939
BLEUscore(J)	0.906	0.912	0.994	1	0.928
PPscore(J)	0.916	0.929	0.939	0.928	1

表 3: 類似度間の相関係数

聞取り能力が高いと、「正解の英語の発話」を見せられても「聞取った英語の発話の和訳文」をあまり修正しなかったということが分かる。また、同じ表より、異なる類似度算出法の間にも高い相関があることが分かる。

4 TOEIC_MAD_EJ データとリファレンスの解析

翻訳一対比較法の自動化手法が英日方向の翻訳評価に応用できるかどうかを明らかにするために、TOEIC_MAD_EJ データとリファレンスを合わせて解析した。リファレンス (複数の正解翻訳) としては、テストセットの発話 1 発話に対して、15 の翻訳を人手で作成したものを用いた。

被験者の翻訳結果とリファレンスの類似度と TOEIC スコアの間に正の相関があれば、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ことができる。

以下、「正解の英語の発話の和訳文」とリファレンスの類似度と TOEIC スコアの関係、および、「聞取った英語の発話の和訳文」とリファレンスの類似度と TOEIC スコアの関係を調査した結果を示す。類似度の計算には、DP マッチングと最大 n-gram 長を 2 とした BLEU を用いた。

4.1 「正解の英語の発話の和訳文」とリファレンスの類似度と TOEIC スコアの関係

「正解の英語の発話の和訳文」とリファレンスの類似度を求め、それと TOEIC スコアとの関係を調査した。結果を図 6に示す。

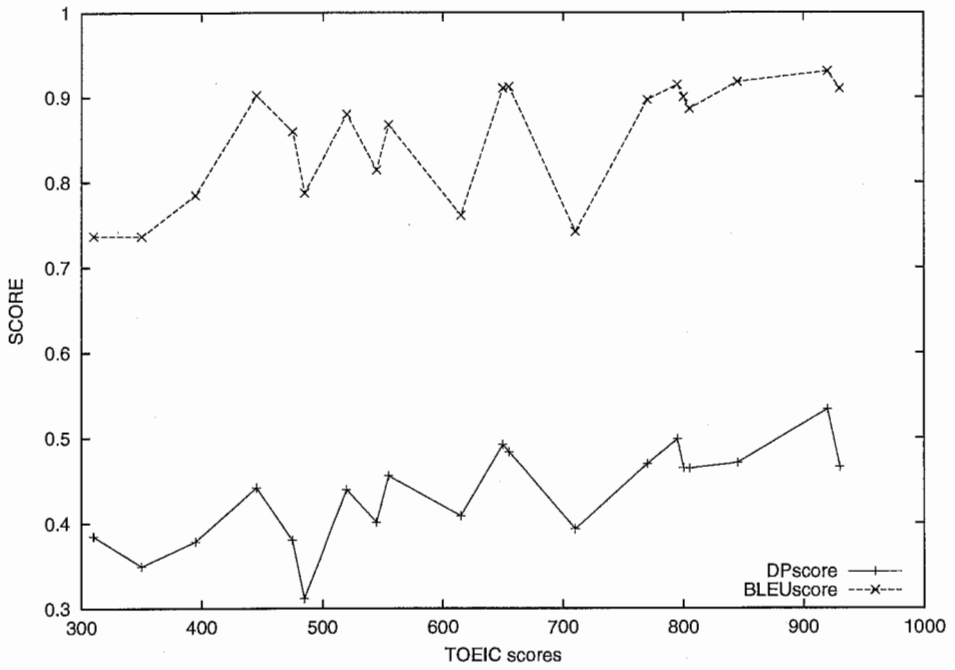

図 6: 「正解の英語の発話の和訳文」とリファレンスの類似度

DP マッチングによる類似度と TOEIC スコアの相関係数は 0.745、BLEU による類似度と TOEIC スコアの相関係数は 0.659 であった。「正解の英語の発話の和訳文」とリファレンスの類似度と TOEIC スコアは相関しているようである。

4.2 「聞取った英語の発話の和訳文」とリファレンスの類似度と TOEIC スコアの関係

「聞取った英語の発話の和訳文」とリファレンスの類似度を求め、それと TOEIC スコアとの関係を調査した。結果を図 7に示す。

DP マッチングによる類似度と TOEIC スコアの相関係数は 0.791、BLEU による類似度と TOEIC スコアの相関係数は 0.752 であった。「聞取った英語の発話の和訳文」とリファレンスの類似度と TOEIC スコアは相関しているようである。

5 英日翻訳システムの評価実験

前章で、それぞれの和訳文に対して、和訳文とリファレンスの類似度と TOEIC スコアが相関していることが分かった。そこで、TOEIC_MAD_EJ データとリファレンスを用いて、翻訳一対比較法の自動化手法を英日方向の翻訳システムの評価に応用する実験を行った。ここでは評価対象として、音声翻訳システム ATR-MATRIX の言語翻訳サブシステム TDMT の MAD1 version を用いた。また、英語の native speaker が発話した英語をタイピストがタイプしたものを TDMT の入力とした。タイピストはほとんどタイプミスをしていないため、「正解の英語の発話」とほぼ等しいテキストが入力されたと考えてよい。

DP マッチングによる、「システムの翻訳」とリファレンスの類似度は 0.346、BLEU による、「システムの翻訳」とリファレンスの類似度は 0.809 であった。これらと、前章で求めた被験者の和訳文のスコアからシステム TOEIC 換算点を求めた。

DP ベース自動評価法を用いて被験者の和訳文のスコアを計算し、回帰直線を引いたものを図 8に示す。BLEU を用いて被験者の和訳文のスコアを計算し、回帰直線を引いたものを図 9に示す。これらの図における水平の直線は、「システムの翻訳」のスコアを表している。

それぞれの場合において、回帰分析によってシステム TOEIC 換算点を求めた。その結果を表 4に示す。ここで、「聞取った英語の発話の和訳文」とリファレンスの類似度を用いてシステム TOEIC 換算点を求めると

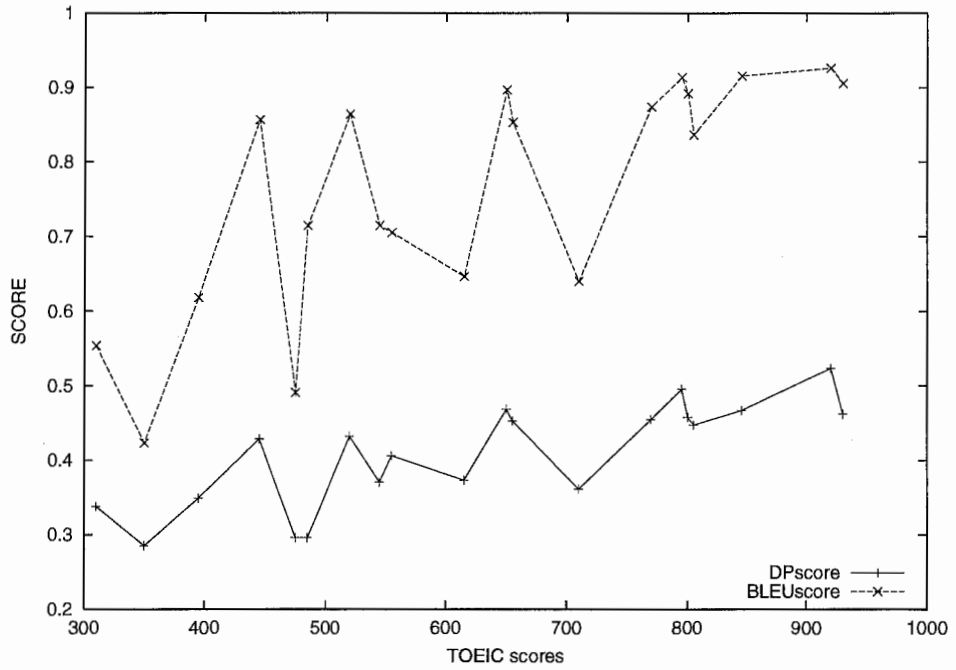

図 7: 「聞取った英語の発話の和訳文」とリファレンスの類似度

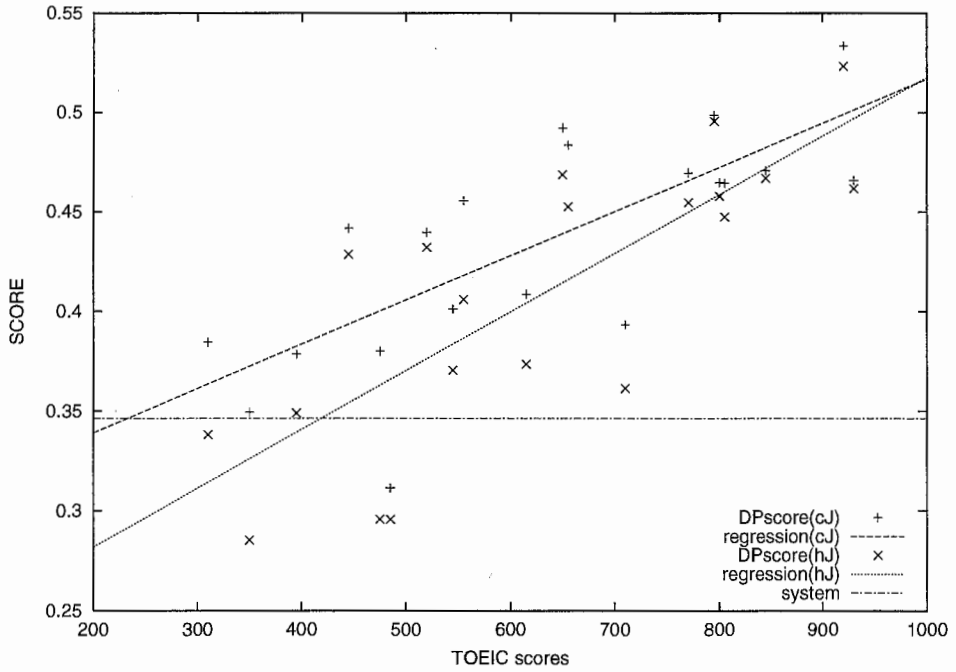

図 8: DP ベース自動評価法による被験者の和訳文のスコアの回帰直線

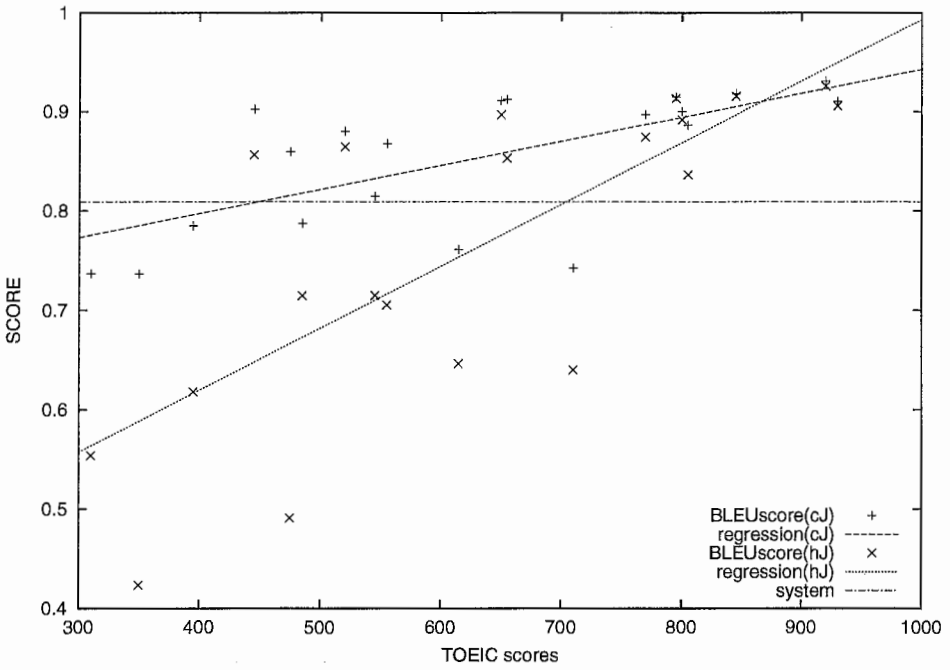

図 9: BLEU による被験者の和訳文のスコアの回帰直線

	DP マッチング	BLEU
「正解の英語の発話の和訳文」とリファレンスの類似度を用いた場合	233	448
「聞取った英語の発話の和訳文」とリファレンスの類似度を用いた場合	419	704

表 4: システム TOEIC 換算点

いうことは、音声翻訳システムの音声認識器の認識精度が 100% という状態におけるシステムの TOEIC スコアを求めることに相当する。

6 種々の考察

前章で求めたシステム TOEIC 換算点は、翻訳自動評価法の選択によってかなり異なった値となった。その原因について考察する。また、新たに考案した翻訳自動評価法を用いて、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した。この翻訳自動評価法について述べ、この手法で求まったシステム TOEIC 換算点を示す。最後に、リファレンス数の影響について考察する。

6.1 前章の結果の考察

「正解の英語の発話の和訳文」とリファレンスの類似度、「聞取った英語の発話の和訳文」とリファレンスの類似度のどちらを用いた場合においても、DP ベース自動評価法を用いて算出したシステム TOEIC 換算点は、BLEU を用いて算出したシステム TOEIC 換算点よりも低い値となった。この原因について考察する。

システムは入力に対して必ずある程度の出力を返す。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。この事実により、システムの BLEU スコアのペナルティーの値は小さくなり、その結果、システムの BLEU スコアは相対的に高い値になったと思われる。

次式が BLEU スコアのペナルティーを与える式である。

$$S_{penalty} = \exp\left\{\max\left(\frac{L^*_{ref}}{L_{sys}} - 1, 0\right)\right\} \tag{4}$$

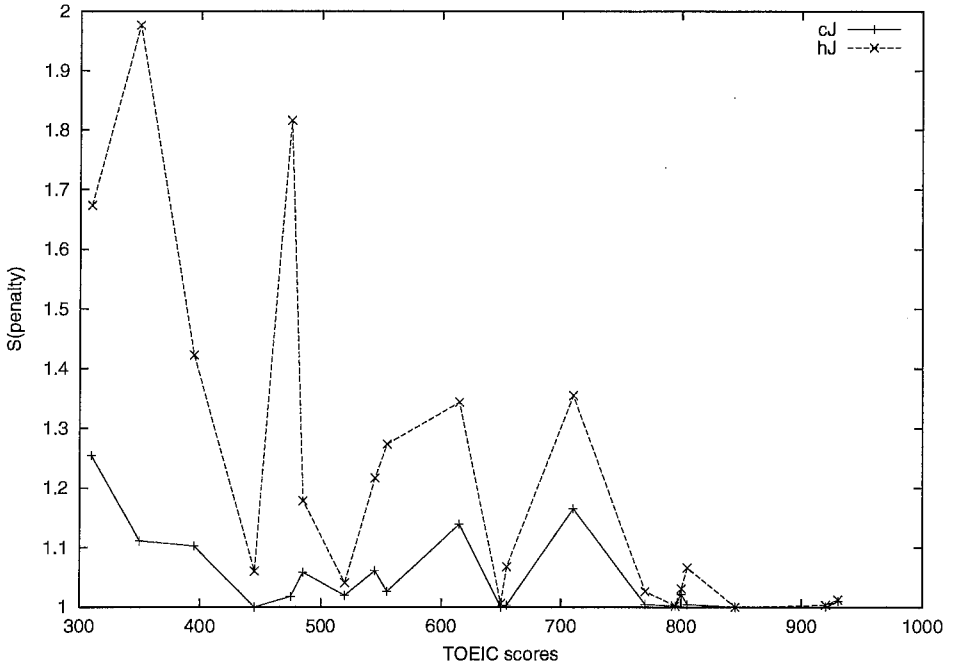

図 10: BLEU スコアのペナルティー

実際に被験者の BLEU スコアの $S_{penalty}$ とシステムの BLEU スコアの $S_{penalty}$ を求めた。「正解の英語の発話の和訳文」の BLEU スコアの $S_{penalty}$ と、「聞取った英語の発話の和訳文」の BLEU スコアの $S_{penalty}$ を図 10に示す。システムの BLEU スコアの $S_{penalty}$ は 1.00 であった。

図 10から、上の推測が正しいことが分かる。

一方、翻訳一対比較法でシステムの評価を行っていないので、算出したシステム TOEIC 換算点の絶対的な値について議論することはできない。どの翻訳自動評価法を用いれば翻訳一対比較法と近い評価結果が得られるかを知るためには、翻訳一対比較法を行う必要がある。

6.2 新たな翻訳自動評価法

ここでは、新たに考案した翻訳自動評価法を二つ述べる。一つめは DP 平均スコアであり、これは DP スコアにおいて max を取るところで平均をとったものである。二つめは DP 類似 Perplexity であり、これは DP マッチングで一番類似度が高い正解翻訳をリファレンスの中から一つ取り出して、その取り出した正解翻訳の集合の Perplexity を求めるというものである。以下、その翻訳自動評価法について詳説し、それを用いて翻訳一対比較法の自動化手法を行って求めたシステム TOEIC 換算点を示す。評価対象としては、先ほどと同じく音声翻訳システム ATR-MATRIX の言語翻訳サブシステム TDMT の MAD1 version を用いた。

6.2.1 DP 平均スコア

DP 平均スコア S_{DPmean} は、以下のように定義される。

$$S_{DPmean} = \frac{1}{N_{total}} \sum_{j=1}^{N_{total}} \frac{1}{N_{ref}} \sum_{i=1}^{N_{ref}} \max\left(\frac{T_{ij} - S_{ij} - I_{ij} - D_{ij}}{T_{ij}}, 0\right) \tag{5}$$

ここで、N_{ref} はリファレンス数を示す。式 (5) は、式 (1) において max を取っていたところを、リファレンスの平均に置き換えた式である。

DP 平均スコアを用いて被験者の和訳文のスコアを計算し、回帰直線を引いたものを図 11に示す。図における水平の直線は、システムのスコアを表している。

回帰分析によって求めたシステム TOEIC 換算点を表 5に示す。

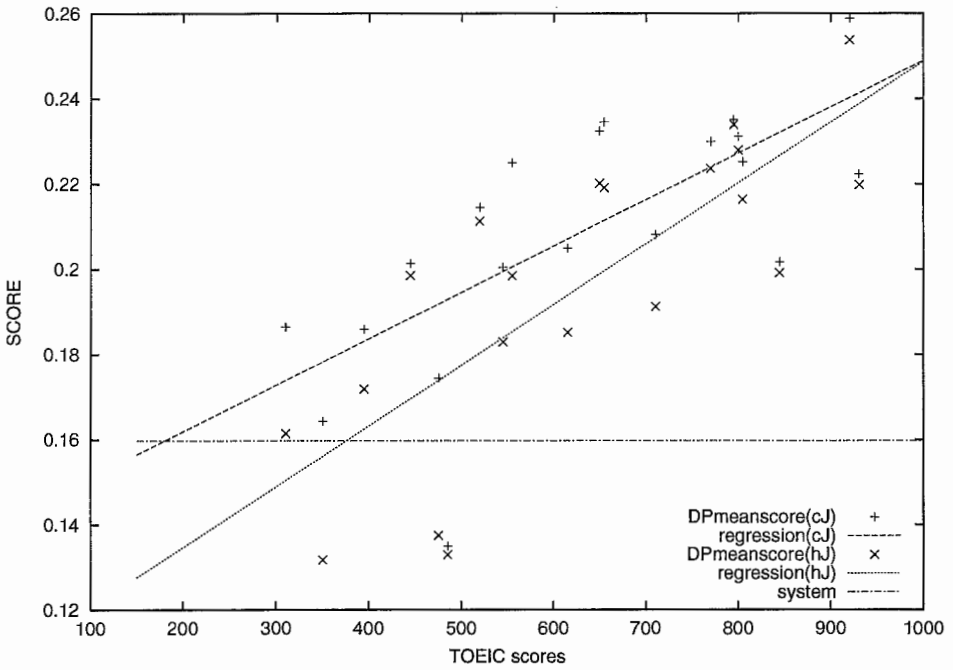

図 11: DP 平均スコアによる、被験者のスコアとその回帰直線、およびシステムのスコア

	システム TOEIC 換算点
「正解の英語の発話の和訳文」とリファレンスの類似度を用いた場合	180
「聞取った英語の発話の和訳文」とリファレンスの類似度を用いた場合	376

表 5: 翻訳自動評価法に DP 平均スコアを選んだ時のシステム TOEIC 換算点

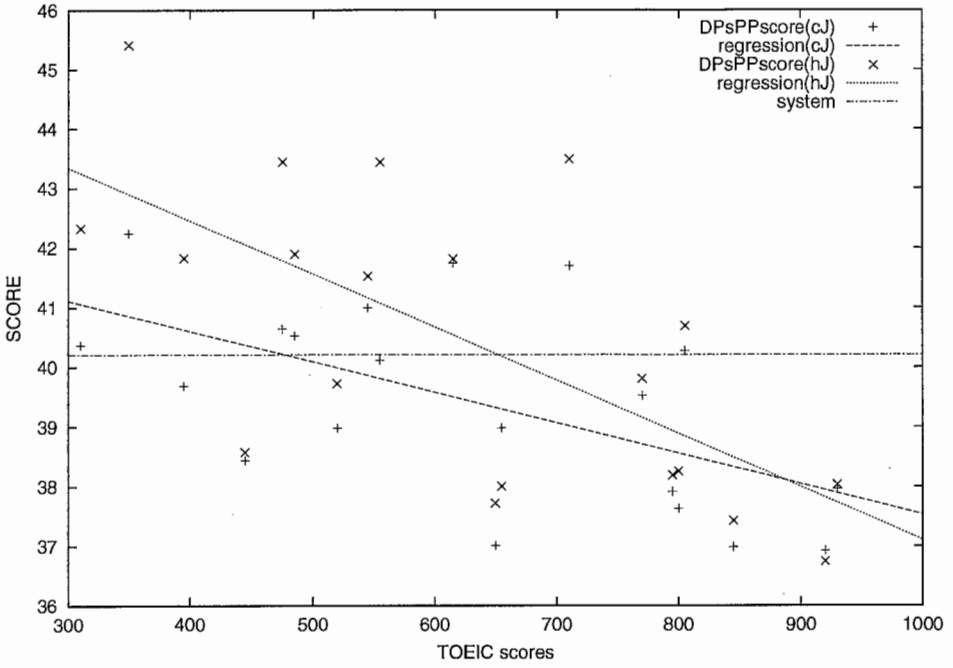

図 12: DP 類似 Perplexity による、被験者のスコアとその回帰直線、およびシステムのスコア

	システム TOEIC 換算点
「正解の英語の発話の和訳文」とリファレンスの類似度を用いた場合	477
「聞取った英語の発話の和訳文」とリファレンスの類似度を用いた場合	652

表 6: 翻訳自動評価法に DP 類似 Perplexity を選んだ時のシステム TOEIC 換算点

6.2.2 DP 類似 Perplexity

テストセットの各々の発話に対して、評価対象の翻訳結果と DP マッチングで一番類似度が高い正解翻訳をリファレンスの中から一つ取り出す。その取り出した正解翻訳の集合の Perplexity を DP 類似 Perplexity と呼ぶことにする。DP 類似 Perplexity は以下の手順で求めることができる。

1. DP マッチングによって、評価対象の翻訳結果とリファレンスのそれぞれの正解翻訳との類似度を求める
2. 一番類似度が高かった正解翻訳を一つ取り出す
3. 上の 1. 2. の操作をすべての発話に対して行う
4. 取り出した正解翻訳の Perplexity を求める

各和訳文に対して DP 類似 Perplexity を計算し、回帰直線を引いたものを図 12に示す。図における水平の直線は、システムの DP 類似 Perplexity を表している。

回帰分析によって求めたシステム TOEIC 換算点を表 6に示す。

6.3 リファレンス数の考察

翻訳自動評価法においてリファレンス数がどれほどあれば十分であるか調べるために、リファレンス数を 1 から 15 まで 1 ずつ変化させながらシステム TOEIC 換算点を求めた。

1 から 15 までの任意の k に対して、15 の正解翻訳の中からランダムで k 個の正解翻訳を選び出してそれをリファレンスとし、そのリファレンスを用いて翻訳一対比較法の自動化手法を行った。翻訳自動評価法には、

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

表 7: リファレンス数に対するシステム TOEIC 換算点

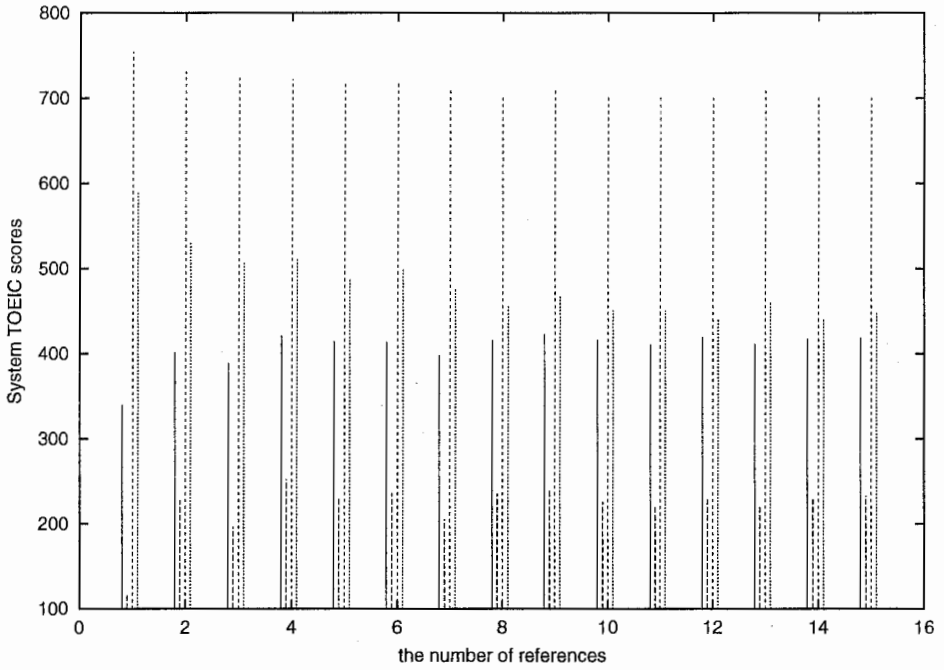

図 13: リファレンス数に対するシステム TOEIC 換算点

DP ベース自動評価法と最大 n-gram 長を 2 とした BLEU を用いた。評価対象としては、先ほどの音声翻訳システム ATR-MATRIX の言語翻訳サブシステム TDMT の MAD1 version を用いた。

得られた結果を表 7に示す。表中の「DP + hJ」は、DP ベース自動評価法によって「聞取った英語の発話の和訳文」のスコアを求め、それを用いてシステム TOEIC 換算点を算出したということを示す。「DP + cJ」は、DP ベース自動評価法によって「正解の英語の発話の和訳文」のスコアを求め、それを用いてシステム TOEIC 換算点を算出したということを示す。「BLEU + hJ」は、BLEU によって「聞取った英語の発話の和訳文」のスコアを求め、それを用いてシステム TOEIC 換算点を算出したということを示す。「BLEU + cJ」は、BLEU によって「正解の英語の発話の和訳文」のスコアを求め、それを用いてシステム TOEIC 換算点を算出したということを示す。

表 7をグラフにしたものが図 13である。棒グラフは左から順に「DP + hJ」、「DP + cJ」、「BLEU + hJ」、「BLEU + cJ」を表している。

表 7と図 13から、リファレンス数が 10 以上になると、リファレンス数の増加におけるシステム TOEIC 換算点の増減はほとんどないということが見て取れる。従って、コストを抑えながらも信頼のあるスコア計算に十分なリファレンスを作成するためには、リファレンス数を 10 程度にすればよいということが分かる。

7 まとめと今後の課題

日英方向の翻訳評価法は、英日方向の翻訳評価に応用できる見込みがあることが分かった。また、翻訳自動評価法で用いるリファレンスのリファレンス数は 10 程度あれば十分であることが分かった。

今回の評価実験で用いた翻訳システムの翻訳結果は対話データ収集の際のものをそのまま利用したため、用いた翻訳自動評価法ではそのシステムに正しくスコアがつけられていない可能性がある。この点を明らかにするために、別の翻訳システムの評価実験を行う必要がある。また、日本語は語順が比較的自由であるため、リファレンスの文の文節を文法的に正しく入れ替えた文を評価対象の翻訳結果とした場合、高いスコアが与えられるべきであるが、DP ベース自動評価法においては低いスコアしか与えられない。この点を考慮した、新たな翻訳自動評価法を考案する必要があるように思われる。

謝辞

本研究を行う機会を与えてくださいました 音声言語コミュニケーション研究所 菊井玄一郎 室長、京都大学情報学研究科 佐藤理史 先生に感謝いたします。また、助言や討論をいただいた 安田圭志 研究員、林輝昭 氏をはじめとする第二研究室の皆様に感謝いたします。

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る。

参考文献

[1] 菅谷史昭：音声翻訳品質の評価に関する研究、神戸大学博士論文 (2003)

[2] TOEIC: Test of English for International Communication(2002). http://www.toeic.com/

[3] 安田圭志、菅谷史昭、竹澤寿幸、山本誠一、柳田益造：表層単語列のパタンから計算される客観尺度を用いた翻訳一対比較法の自動化, 電子情報通信学会論文誌 (投稿中)

[4] Papineni, K.,Roukos, S.,Ward, T. and Zhu, W.-J.: Bleu:a Method for Automatic Evaluation of Machine Translation , *Proc. ACL*, pp.311-318(2002).

A 付録

A.1 Perplexity

Perplexity PP は、次式によって定義される。

$$\log_2 PP = -\frac{1}{L}\sum_{i=1}^{L}\log P(w_i|w_{i-N+1}^{i-1}) \tag{6}$$

つまり、Perplexity PP は、H をエントロピーとするとき、$PP = 2^H$ によって定義される。L として単語数を用いると、word Perplexity となり、L として文の数を用いると、sentence Perplexity となる。

A.1.1 Perplexity の比 (エントロピーの差)

Perplexity の比 S_{PP} は、以下のように定義される。

$$S_{PP} = \frac{PP_{sys}}{PP_{ref}} \tag{7}$$

ただし、PP_{sys} は評価対象の翻訳の Perplexity を、PP_{ref} はリファレンスの Perplexity を示す。式 (7) は、評価対象の翻訳のエントロピーを H_{sys}、リファレンスのエントロピーを H_{ref} とすると、次のように書き換えられる。

$$\begin{aligned} S_{PP} &= \frac{PP_{sys}}{PP_{ref}} \\ &= \frac{2^{H_{sys}}}{2^{H_{ref}}} \\ &= 2^{H_{sys}-H_{ref}} \end{aligned} \tag{8}$$

式 (8) の右辺は、2 の肩に「エントロピーの差」が乗っている形をしている。このことから、上で定義した Perplexity の比は、エントロピーの差と呼ぶこともできる。

A.1.2 3.2.2 の補足

3.2.2 の Perplexity の比を計算するときに用いたのは、word Perplexity である。word Perplexity の代わりに sentence Perplexity を用いて Perplexity の比を計算すると、図 14のようになる。このとき、TOEIC スコアとの相関係数は 0.753 である。

sentence Perplexity による Perplexity の比は、DP マッチングによる類似度、BLEU による類似度との相関が相対的に低い (それぞれ 0.703、0.691) ため、本文中では word Perplexity による Perplexity の比を用いた。

A.2 相関係数

変数 x と y の相関係数 r (Pearson correlation coefficient) は、次式によって定義される。

$$r = \sum_i \frac{(x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y} \tag{9}$$

ここで、n は測定数、y_i は $x = x_i$ の時の y の実測値、$\bar{x}, \bar{y}$ はそれぞれ x, y の平均値、σ_x, σ_y はそれぞれ x, y の標準偏差を示す。

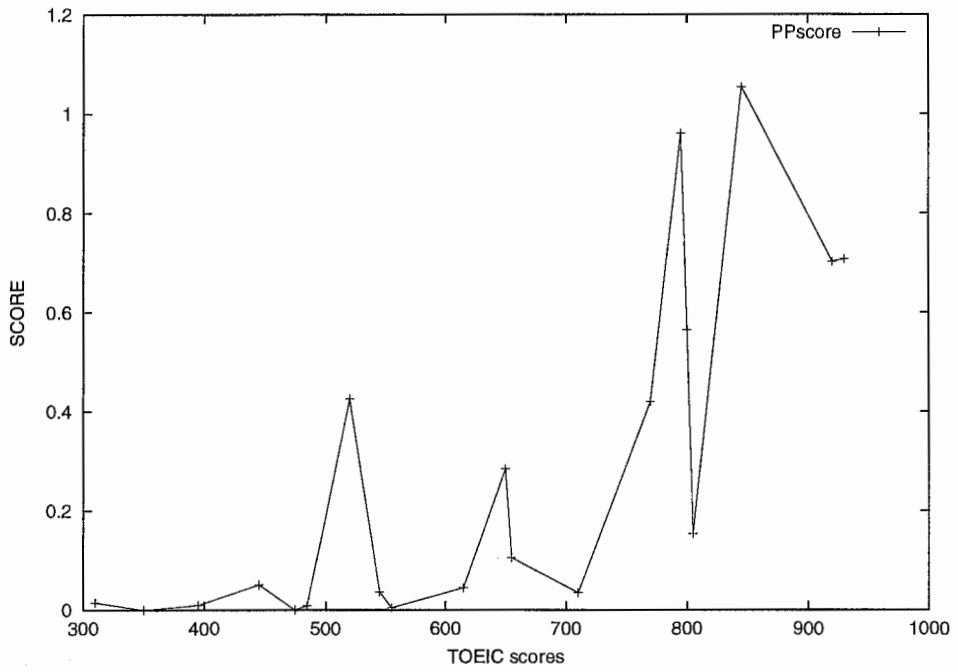

図 14: sentence Perplexity による
「正解の英語の発話の和訳文」と「聞取った英 語の発話の和訳文」の Perplexity の比