機械翻訳システムの有効性の評価

～どのような人にとってMTは役立つか～

富士秀（富士通研）、畠中伸敏（キヤノン）、伊藤悦雄（東芝）、亀井真一郎（NEC）、
限井裕之（日立）、介弘達也（沖）、吉見毅彦（シャープ）、井佐原均（通信総研）

fuji.masaru@jp.fujitsu.com

1. はじめに

インターネットの普及にともなって、外国語、特に英語のウェブページに直接アクセスする機会が増えている。このような状況で機械翻訳は、外国語文書を母語に変換し、斜め読みするためのツールとして使われ始めている。現状の機械翻訳システムは、まだまだ「完璧な」翻訳を行なうことはできないため、たとえば、英語力のある人は、英日機械翻訳の出力よりは英語を直接読んだほうがいいと考えているだろう。しかし、同時に英語に自信がなく、完璧でなくても何かに頼りたいと思う多くの人々がいることも確かであろう。本研究の目的は、現状の機械翻訳システムがどのようなユーザ層にとって有益であるかを見積ることである。

なお本研究は、アジア太平洋機械翻訳協会（AAMT）技術動向調査委員会における調査活動の一環として行われた。協会の業界団体的性格から、研究者や開発者ではない一般の人にとって直感的な結果の出る評価手法の設計を目指した。

2. 評価手法

今回の実験では、機械翻訳システムの品質の尺度として、多くの受験者を持つ TOEIC を用いた。ここでの考え方を一言でいえば、TOEIC で 500 点を取った人が機械翻訳の出力を利用することによって 600 点の成績を修めることができれば、この人にとってはこの機械翻訳システムは有効であるといえる、ということである。

我々はユーザが機械翻訳の出力を有益であるかどうか決めるのは、その人の英語理解力に依存するであろうと考え、これを確認するために、被験者に TOEIC の読解文を、原文（英語）単独、原文の機械翻訳文（日本語）単独、これらの両者併用（英語および日本語）の３形態で回答してもらった。同時に、上記３形態のどれを使った場合がもっとも回答が容易であったかという印象についてアンケートに答えてもらった。

なお、機械翻訳全般の従来評価手法としては、①テストセットを用いるもの[1][2]、②熟練評価者による主観評価[3]、③（読解問題等の）タスクにおける成績による評価[4][5][6]、等がある。本手法は③に属するが、被験者の外国語能力との関係の測定や、機械翻訳文と原文の併用時の効果測定等は本論文の特徴である。

3. 実験の仕様

TOEIC テストの利用部分

オリジナルの TOEIC テストは、大きく、リスニングとリーディングの２つのセクションに分かれている。リーディングセクションはさらに３パートに分かれていて、この最後のパートが読解問題であるが、本評価実験ではこの読解問題の部分のみを使う。リーディングセクション全体は 100 問の選択問題で構成されており、全体を 75 分で回答することになっている。このうちの読解問題には 40 問の選択問題がある。

実験パターン

機械翻訳の利用方法としては、訳文を単体で読む場合と、訳文と原文とを並べて読む場合があるため、今回の実験では、その両方をテストした。2 種類の機械翻訳システム（MT1 および MT2）を使用したため、各被験者は、英語原文（1 種類）、機械翻訳文（2 種類）、両者併用（2 種類）の５つのパターンのテストに答えることになる。毎回新しい文を被験者に提示する必要があるため、各パターンは別の問題から作成される。実験パターンを表 1 にまとめる。

表 1：実験パターンの一覧

パターン 1	原文および質問文（英語）
パターン 2	原文と質問文の MT 1 による和訳（日本語）
パターン 3	原文（英語）および原文と質問文の MT 1 による和訳（日本語）を並べて提示
パターン 4	原文と質問文の MT 2 による和訳（日本語）
パターン 5	原文（英語）および原文と質問文の MT 2 による和訳（日本語）を並べて提示

時間設定

実際の TOEIC テストと同じような環境で実験を行なうため、1 セットの読解問題の回答時間は 40 分とした。これはリーディングセクション（100 問 75 分）における読解問題（40 問）の比率をもとに、読解問題が他の問題（文法・語彙問題と誤文訂正問題）よりも時間がかかる傾向にあることから定めた。被験者は 5 つのパターンに答えるため、各被験者の所要時間は 200 分となる。

なお、各パターンの中では、被験者は回答が終了した時点で終了時刻を記入して退室してもよいこととした。残り時間の使い方に個人差があるため指標として若干の問題はあるが、この回答時間もある程度の意味を持つと考えて分析を行なった。

読解テスト後のアンケート

今回の実験に参加した全被験者は、読解テスト終了後に、読解文書に対する印象についてのアンケートに答えてもらうようにした。読解得点や回答時間はある程度機械的に定量測定できるものだが、システムの有益性を検証するにはこのような印象も重要だと考えたからである。

機械翻訳システム

実験用の機械翻訳テキストを得るために、2 種類の商用機械翻訳システム（MT 1 および MT 2）を利用した。これは結果がシステムの差異によって左右されるかどうかを確認するためである。

被験者

今回の実験では、さまざまな英語能力の被験者を集めることが必要になる。200 名の被験者は TOEIC のスコアが分散するように集められた。（表 2）

表 2: 被験者グループと人数

グループ名	得点	人数
G1	10-390	5
G2	395-440	10
G3	445-490	18
G4	495-540	19
G5	545-590	19
G6	595-640	19
G7	645-690	19
G8	695-750	23
G9	745-790	18
G10	795-840	11
G11	845-890	10
G12	895-	12

4. 実験の条件

機械翻訳のチューニング

機械翻訳システムは、単語登録などによってチューニングすることが可能であり、これは訳質に大きく影響する。今回の実験では、チューニングのレベルをコントロールすることが困難であるため、チューニングはまったく行なわないことにした。

機械翻訳の実行モード

文書のフォーマットをそのままに翻訳できなければ、読みやすさは激減する。今回の実験では読解用文章は HTML 化し、機械翻訳システムの WWW ブラウザモードで翻訳した。

設問の翻訳

TOEIC の読解問題は、読解すべき文書と理解度を計るための設問からなっている。設問を機械翻訳システムで翻訳するのか、人手によって（正しく）翻訳するのかという選択肢がある。人手による翻訳の場合、システムの訳語と一致しないなどの問題が起こる可能性があり、今回の実験では手順の単純化のため、設問も同じ機械翻訳システムによって翻訳することにした。

5. 結果と分析

機械翻訳システムによる差異

実験に用いた二つの機械翻訳システムの出力する訳文は、一つ一つ比較すれば、ある程度の異なりが生じている。これが今回の実験において有意な差異となるかどうかを被験者グループ毎に T 検定で検証した。

この結果、機械翻訳文を単独で用いた場合も、機械翻訳文と原文を併用した場合も、システムの異なりによる有意な差異は認められなかった。よって、以下の議論では 2 つのシステムの結果を合わせて議論する。

読解得点

原文の読解（基準値）

実験の基準値として使うために、原文に対する読解実験を行なった（表 1 の「パターン 1」）が、この結果を図 1 に示す。図中、横軸には表 2 のグループ名を、縦軸に今回の読解問題の得点をプロットした。全問正解すると 40 点となる。

被験者の申告した過去の TOEIC 点数は、今回実施する読解問題のみの点数と強い関係があることが予想されるが、これを裏付ける結果となっている。

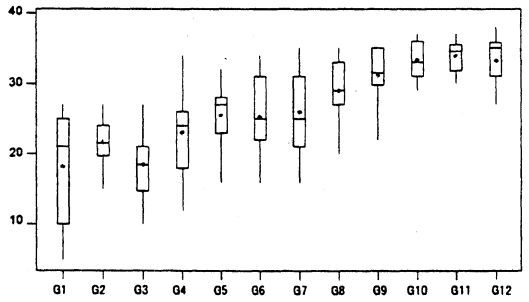

図 1: 原文のみの提示による読解得点

機械翻訳文単独の読解

図 2 は、機械翻訳出力の和文のみを被験者に提示した場合の被験者グループ毎の読解得点を表している。図 1 の原文のみの結果と比較すると、TOEIC 低得点層では読解得点にほとんど差が見られないが、高得点層では和文を提示したほうがむしろ読解得点が低くなっている。このことから、機械翻訳文のみの提示は、TOEIC 低得点層では効果が認められず、高得点層ではむしろ読解の妨げとなっている。

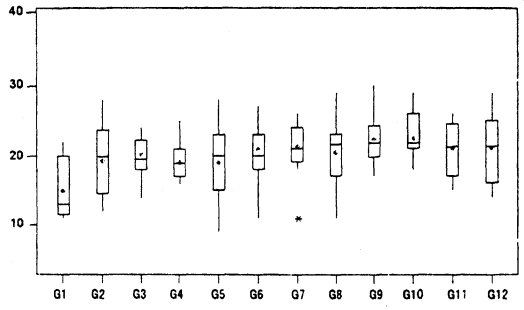

図 2：機械翻訳文単体での得点

機械翻訳文と原文の併用

図 3 は、機械翻訳和文と原文の両方を同時に被験者に提示したときの読解得点である。図 1 と比較すると、TOEIC 高得点層では読解得点に差がないが、低得点層では図 3 の方が読解点数が高い傾向にある。このことから、機械翻訳文と原文を併用すると、TOEIC 高得点層での読解低下を招くことなく、低得点層での読解得点改善が起こることがわかる。

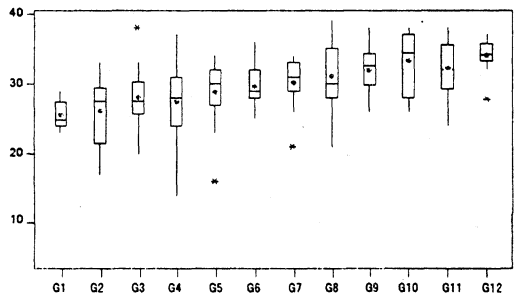

図 3：機械翻訳和文と原文の併用

有意性の検定

上述の読解得点に関する実験結果の検定を行なった。表 3 では、原文のみの結果と比較したときの検定結果を掲載している。ここでは T 検定の値が 0.1 以下の場合を「有意」と見なしている。

「MT 単独」とは、原文のみの結果と機械翻訳和文単独の結果の間の検定結果である。ここでは、G3 および G8～G12 で有意差が見られる。前項から、この有意差は「原文のみの方が読解得点が高い」という方向における有意差である。

「MT＋原文」とは、原文のみの結果と、機械翻訳和文＋原文併用時の結果の間の検定結果である。G1～G7 で有意差が検出されたが、これは前項から、「原文のみよりも併用の方が得点が高い」という方向での有意差である。

表 3：T 検定よる原文単独時の得点との比較

グループ	MT単独	MT＋原文
G1	0.4975	0.0958
G2	0.5053	0.0271
G3	0.0013	0.0000
G4	0.4900	0.0051
G5	0.1426	0.0143
G6	0.2715	0.0026
G7	0.2298	0.0087
G8	0.0021	0.7129
G9	0.0003	0.1449
G10	0.0008	0.9255
G11	0.0003	0.4778
G12	0.0000	0.6550

回答所要時間

図 4 は、各パターンの平均回答所要時間を示している。ここでも原文のみの際の読解得点との比較の議論を行なう。読解得点と同様に、TOEIC 高得点層では有意差が見られないが、低得点層では機械翻訳と原文の併用の効果が出ている。

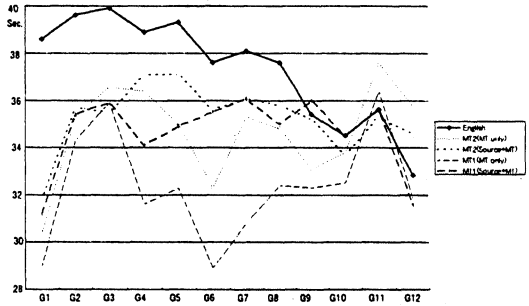

図 4: 平均回答所要時間

表 4 は、平均所要回答時間についての、原文のみの時間を基準とした T 検定の結果である。ここでも、TOEIC 低得点層での機械翻訳と原文併用パターンによる効果が現れている。ただし回答所要時間は他の要因にも影響されているので議論はこれに留める。

表 4：回答所要時間の有意差検定

	MT 単独	MT＋原文
G1	0.0276	0.1021
G2	0.0027	0.0016
G3	0.0005	0.0008
G4	0.0002	0.0001
G5	0.0000	0.0009
G6	0.0000	0.0394
G7	0.0004	0.0469
G8	0.0004	0.0301
G9	0.0078	0.9432
G10	0.5897	0.1087
G11	0.5199	0.3486
G12	0.8563	0.4769

分かりやすさに関する被験者の印象

図 5 は、機械翻訳文のみの場合と、機械翻訳文と原文併用の場合について、「分かりやすさ」に関する印象を被験者アンケートから得た結果である。分かりやすさでは、「分かりやすい」が 5 点、「分かりにくい」が 1 点であるような 5 段階評価をつけてもらった。
表 5 では、中点である「評価 3」を基準に分析する。この結果、読解得点と同様の傾向が現れており、併用パターンにおいて TOEIC 低得点層で「分かりやすい」と感じる被験者が多い。

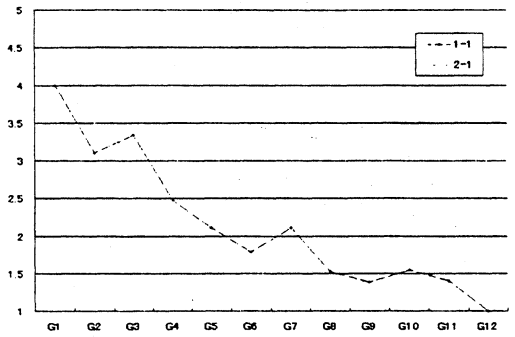

図 5: 分かりやすさ

1-1　MT のみの分かりやすさ
2-1　MT＋原文の分かりやすさ

表 5 は、提示された文章に対する「分かりやすさ」の T 検定結果である。中点である得点 3 を基準とした検定となっている。
結果は読解得点と同様の傾向を示しており、TOEIC 高得点層では機械翻訳文単独は分かりやすさが有意に低下しており、低得点層では併用パターンで分かりやすさが有意に向上している。ただし、有意差が認められるのが G1～G3 と、読解得点の時と比べてより狭い得点層（特に低い得点層）に限られていることがわかる。

表 5: 分かりやすさの検定

	MT 単独	MT＋原文
G1	0.8193	0.9995
G2	0.5382	0.6610
G3	0.6136	0.9088
G4	0.3223	0.6442
G5	0.1468	0.5741
G6	0.0571	0.6610
G7	0.1776	0.6110
G8	0.1009	0.6236
G9	0.0165	0.4013
G10	0.0015	0.4429
G11	0.0005	0.3597
G12	0.0000	0.2989

6. 結論

本評価手法では、ある与えられた環境（機械翻訳システム、システムのチューニングレベル、被験者、等）において、機械翻訳の有効性に関する統計的に有意な結果を得ることができた。結果は、読解得点・回答所要時間・有益性の印象という 3 つの異なる軸で得られた。今回の環境における実験結果についてまとめる。

- **機械翻訳文単独 vs. 原文との併用**
  機械翻訳文単独の利用と、機械翻訳文と原文の併用による利用では、いずれの評価軸でも同様の傾向が見られた。単独利用による有益性（読解得点向上、読解時間短縮、分かりやすさ向上）はどのユーザ層にも見られず、TOEIC 高得点者ではむしろ機械翻訳の有益性が減少する。これに対して、併用利用では TOEIC 低得点層で有益性が向上する。
- **読解得点と分かりやすさ**
  読解得点では G1～G7 という広い低得点層で得点が増加するが、分かりやすさの印象という軸では、G1～G3 というより狭い低得点層でのみ有意となっている。

以上の議論をまとめると、今回の実験環境での結論は次のようになる。

TOEIC 得点	読解得点向上	分かりやすさ向上
～490	○	○
495～690	○	×
690～	×	×

謝辞

本研究における評価実験の準備、遂行、データ整理等の一連の作業をお願いした、柴田葉子氏・大野悟氏をはじめとする㈱アイアール・アルトの方々に感謝いたします。

参考文献

[1] Hitoshi Isahara, et al., (1995) JEIDA's Test-Sets for Quality Evaluation of MT Systems. In proceedings of MT-Summit V.

[2] Sungryong Koh, Jinee Maeng, Ji-Young Lee, Young-Sook Chae, Key-Sun Choi (2001). A Test Suite for Evaluation of English-to-Korean Machine Translation Systems. In proceedings of MT Summit, pp.191—195.

[3] Maki Darwin, (2001). Trial and Error: An Evaluation Project on Japanese <-> English MT Output Quality. In proceedings of MT Summit, pp.77—82.

[4] Tomita M., Shirai, M., Tsutsumi, J., Matsumura, M. and Yoshikawa, Y.(1993). Evaluation of MT Systems by TOEFL. In proceedings of TMI, pp.252—259.

[5] Masaru Fuji, (1999). Evaluation Experiment for Reading Comprehension of Machine Translation Outputs. In proceedings of MT Summit VII.

[6] Fumiaki Sugaya, Keiji Yasuda, Toshiyuki Takezawa and Seiichi Yamamoto (2001). Precise Measurement Method of a Speech Translation System's Capability with a Paired Comparison Method between the System and Humans. In proceedings of MT Summit, pp.345—350.