Vol.2014-NL-219 No.4
2014/12/16

中学生は機械翻訳された英語対話文完成問題を解けるか？

藤田 彬† 松崎拓也‡ 登藤直弥† 新井紀子†

機械翻訳器の日常会話翻訳に対する性能を評価する試みについて紹介する．前もって日本語に機械翻訳された英語の対話文完成問題を被験者が解き，その得点を用いて翻訳器の外的な評価を行った．300 名超の被験者を集めた大規模な調査により，評価対象とした翻訳器のうち一つが「人間が文脈を考慮せずに行った翻訳」と同等の性能を有することが明らかになった．本発表では，この調査の内容及び結果を詳述するに加え，一般的に用いられる内的な評価（自動・手動）との比較結果を紹介する．また，人間の対話理解において重大な障害をもたらす翻訳誤りとそうでない誤りを分類し，定量的に分析した結果を報告する．

1. はじめに

機械翻訳は，深い言語知識に限らず，推論能力，常識的知識，世界知識，読み手／聞き手の心理状態モデルなど，あらゆる知能を必要とする典型的な AI 完全問題である．Bar-Hillel[1]が「完全自動の高品質翻訳器の構築を目指すことの不合理性」を主張したが，少なくとも英語と日本語のような隔たりのある言語間の機械翻訳については，未だその主張は妥当であるといえる．

現況において適切な方策は，人間と機械翻訳システムの生産的な協働手法を追求することと考えられる．先行研究として，機械翻訳技術を用いて人間の翻訳者の能力を増強し，高い生産性を持った翻訳を実現する，”translator’s amanuensis [7]”の開発が挙げられる．他の方向性を持った研究として，高品質ではなくとも mono-lingual なエンドユーザの要求を十分に満たす機械翻訳を実現する手法を追求する研究が挙げられる．本稿では，後者の方針に沿い，「人間が機械翻訳器の補助を受けて，第二言語能力の試験問題を解く」という状況を想定した基礎的な調査の結果を報告する．

主に，外国語の対話に関する理解度を測定する目的で設計された典型的な問題形式を取り上げる（図 1 参照）．被験者は，空欄を含む対話を与えられ，選択肢の中から対話中の空欄を埋めるのに最も適切な発話を選ぶ．対話，選択肢の発話は，それぞれ予め機械翻訳される．機械翻訳の補助を受けた対話の実現という目標において現実的なタスクといい難いが，使用される問題は外国語対話の理解度を測定できるように非常に客観的な観点から注意深く設計される．この点で，人間と機械翻訳が協働する複雑な状況における能力の評価・分析に価値のある標本として十分に役立つと考えられる．

これらの問題についての被験者の正答率に基づき、機械翻訳および人手翻訳の両方を含む計 4 種の機械翻訳手法（システム）の評価を行う．評価結果の分析を通じて，現行の機械翻訳の性能とその限界を評価し、どのタイプの翻訳誤りが被験者の補完能力を超えた誤りとなりうるか，または対話の理解を決定的に阻害するか，分析する．

さらに，翻訳文についての外的評価指標（被験者の正答率）と，内的評価指標の関係性を分析する．内的評価指標には，BLEU 値等の一般的に用いられる自動評価と，人手による翻訳自体の質の評価を取り上げる．分析の結果，翻訳された対話の解釈の容易さを直接評価する外的評価指標と内的評価指標の間に相違があることが明らかとなった．

2. 調査手法

2.1 調査の概要

第二言語能力テストから抽出された対話文完成問題についての 4 種の異なる翻訳を，外的に評価した．問題は英語で記述されており，4 種の翻訳手法によってこれらを予め日本語に訳した．2 種の翻訳手法は機械翻訳，別の 2 種は文脈を考慮せずにまたは考慮して人手で訳す手法である．被験者は，どの問題がどの翻訳手法で訳されたものであるかを知らされずに問題を解いた．最後に，同じ翻訳手法で訳された一つの問題を解いた被験者のうち，何名が正解したか（以下，正答率）を指標として，各翻訳手法を評価した．

2.2 被験者

320 名の日本の中学生を被験者とした．被験者は 2 校の生徒から構成される（以下，学校 A，学校 B）．学校 A に所属する被験者は 238 名，残り 82 名は学校 B に所属する．学校 A に所属する被験者は 1 年生が 80 名, 2 年生が 80 名，3 年生が 78 名である．学校 B に所属する被験者は全て 1 年生である．英語の学習状況は学年によって異なり，学校間には学力の差がある．この差が正答率に及ぼす影響については後述する．

2.3 調査資料

大学入試センター試験の模擬試験aから 40 問の英語対話文完成問題をランダム抽出しb，調査資料として用いた．図

† 国立情報学研究所 National Institute of Informatics
‡ 名古屋大学 Nagoya University

a) 大学受験予備校『代々木ゼミナール』のセンター試験模試過去問題集より抽出した．
b) 大学入試センター試験には通例英語対話文完成問題が含まれる．

1 に対話文完成問題の例を示す．全ての問題は，2 名の話者の間で交わされるおよそ 3〜6 文の短い発話（会話部）と，会話部に挿入される可能性のある 4 つの発話（選択肢）で構成される．会話部の一部の発話が空白として隠されており（[BLANK]），被験者はこの空白に当てはまる最も適当な選択肢を選ぶ．全 40 問の問題は，会話部，選択肢全体で 327 文を含む．

2.4 翻訳手法（システム）

下記の 4 種の英日翻訳手法により，英語対話文完成問題を日本語に翻訳したc．Google 翻訳は統計的機械翻訳を，Yahoo 翻訳はルールベース機械翻訳をそれぞれベースとしたシステムと思われる．

1. 機械翻訳システム『Google 翻訳d』 (Set G)
2. 機械翻訳システム『Yahoo 翻訳e』(Set Y)
3. 会話部・選択肢の文順序をランダムに並べ替えて提示した場合の人手翻訳 (Set S)
4. 会話部・選択肢を初期の並びのまま提示した場合の人手翻訳 (Set O)

Set S は，40 問の会話部及び選択肢に含まれる文全てを含むファイルを準備し，文順序を無作為に並び替え，特定の文脈を仮定せずに人手で翻訳をし，元の文順序に再度並べ替えたものである．Set O は，並び替えをせずに文脈を考慮して翻訳したものである．Set S と Set O の翻訳者は同一であり，Set S を訳した後に Set O を訳した．

Set S は，文脈に関する情報を考慮せずに訳される人手翻訳の結果であり，現行機械翻訳システムの性能の上限と実質的に同等の翻訳結果と捉えることができる。

翻訳の専門家ではないが英語が十分に流暢な 3 名の日本語母語話者が，Set S および Set O を個別に作成した．3 セットのうちから 1 セット(同一翻訳者による Set S と Set O の組)を無作為に選出し，調査で被験者に出題する問題とした．残りの 2 セットは，自動評価の参照訳として用いた．

2.5 手順

被験者は 4 つの発話選択肢の中から，会話の流れにおいて最も適当な選択肢を選ぶことで，問題を解いた．被験者は，解答と共に，解答に対する確信の度合（以下，確信度）を下記の 3 段階で示す．

- Level A: 自信をもって答えられた。
- Level B: 他の選択肢と迷った。少し自信がない。
- Level C: 答えを選ぶ決め手が分からない。まったく自信がない。

各被験者には，12 問の問題を出題した．各問の解答時間は，1 分以内に制限した． 12 問には，G, Y, S, O による訳が各 3 問ずつ含まれる．被験者には，「一部の問題が機械翻訳されたもの」であることを事前に告知した．問題の組み合わせは被験者毎に異なる．全 40 問に 4 種類の翻訳結果があるが，この 160 種それぞれの解答者がおおよそ同数ずつとなるように，出題する問題の組み合わせを調整した．

INSTRUCTION
次の会話の BLANK に入れるのに最も適当なものを，それぞれ下の１～４のうちから一つずつ選べ。
DIALOGUE
Receptionist:** Hello. Can I help you?* ***Customer:** Yes. **[BLANK] ***Receptionist:** I'm sorry, I can't find that name on the reservation list.* ***Customer:** Oh, really? Then give me a new reservation, please.*
OPTIONS
1 *I'd like to make a reservation for Flight 502.* 2 *I have a reservation under the name Hashimoto.* 3 *I'm sure you can find my name on the list.* 4 *I wonder if you could tell me how to make a reservation.*

図 1: 多肢選択式対話文完成問題の例 (正答: 2)

2.6 自動評価尺度

被験者の正答率による翻訳の外的評価に加え，いくつかの自動評価尺度による内的評価を行い，外的評価の結果と比較した．

5 種の自動評価尺度 (“BLEU4[10],” “BLEU3[10],” “BLEU+1[9],” “RIBES[6],” および “TER[13]”) に基づき 4 翻訳システムの翻訳結果を評価した．参照訳には，2.4 節に述べた通り，Set S と Set O の組を 2 種用意し，用いた．RIBES と TER の値については，最も良い評価値が出る参照訳セットを用いて計算した値を用いた．

3. 調査結果と考察

3.1 予備調査

学校 A の被験者を学年別に 3 グループに分けて学年間で各問の正答率を比較した．また，学校 A と学校 B の 1 年生の被験者の各問の正答率を比較した．学年及び学校の異なりが正答率に及ぼす効果について，ANOVA により問題別に解析したところ，40 問中 38 問について，学年または学校の異なりそれぞれが正答率に及ぼす影響が有意でなかった($p < 0.05$)．このことから，被験者の学年と学力（英語の能力を含む）が本調査における問題解答に影響しないことがわかる．

3.2 調査結果の概要

4 種の翻訳手法それぞれの評価結果を示す．ここでいう評価結果とは，40 問の正答率と被験者の確信度，自動評価の結果を指す．

図 2 に，40 問の正答率の最小値，最大値，四分位値，中央値を，システム別に箱ひげ図として示す．システム G, Y, S, O の正答率の平均値はそれぞれ，0.524, 0.696, 0.694, 0.875

c) 機械翻訳については，2014 年 6 月 11 日に各サイトで翻訳を実行した．
d) https://translate.google.co.jp/?hl=ja
e) http://honyaku.yahoo.co.jp/

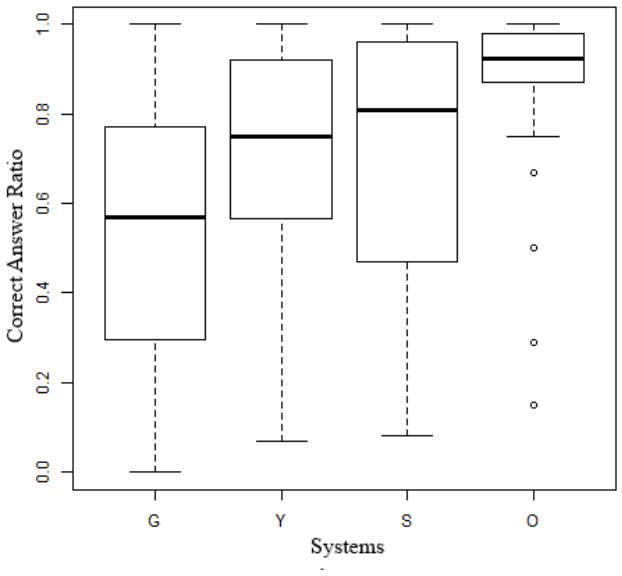

図 2: 40 問の正答率の箱ひげ図

である．システム G-Y 間，Y-S 間，S-O 間について，正答率の差の検定（t 検定）を行ったところ，G-Y 間, S-O 間に有意な差がみられた($p < 0.05$)．一方，Y-S 間には，有意な差が見られなかった($p = 0.954$)．このことから，システム別正答率の大小関係は G < Y ≒ S < O となる傾向にあることがわかる．

図 3 に，システムごとの確信度の分布を示す．Level A がマークされる問題数については G < Y < S < O，Level C については G > Y > S > O となる傾向にあることがわかる．確信度の相対頻度について，問題別にカイ二乗検定を行ったところ, 40 問中 36 問について有意な差がみられた($p < 0.05$)．このことから，翻訳の解釈に関する確信度には 4 システム間で統計的に有意な差があることがわかる．

表 1 に，2 種の参照訳に対する各翻訳システムの自動評価値（5 種）を示す．最下行は，各翻訳システムの正答率の平均である．この結果より，下記のことがわかる．

1) 正答率について G より Y の方が有意に高い一方，BLEU 系評価値及び TER 値は，Y より G の方が高い傾向にある．
2) 正答率について Y と S の間でおおよそ同じである一方, 全ての自動評価値で Y と S の間に大きな差がみられる．
3) 参照訳 S を用いた場合と参照訳 O を用いた場合で自動評価値を比較すると，G, Y, S の評価値は参照訳 S を用いて計算した際により高い値を示す．G, Y, S が文脈を考慮しない人手訳に似た訳であることがわかる．

3)については，現行の機械翻訳器及び Set S の翻訳方針が前後関係を無視して文単位で訳すものであることを考えれば，とりたてて驚くべきものではない．しかしながら少なくとも，S と O の間で平均正答率に有意かつ大きな差が存在するということは，文脈を無視した訳が日常会話の翻訳において大きく解釈性を損ねることを明示する．

次節以降では，これらの結果に関するより微視的な分析の結果について述べる．

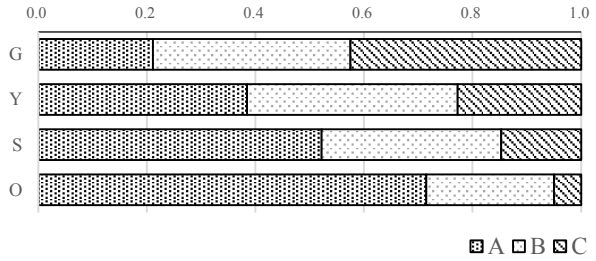

図 3: 翻訳システム別の確信度の分布

表 1: 各参照訳に対する翻訳システム別自動評価値
及び 翻訳システム別平均正答率

Reference	Metrics	G	Y	S	O
Reference O	BLEU4	22.04	20.33	40.30	47.43
	BLEU3	29.09	27.63	47.79	55.09
	BLEU+1	22.08	20.37	40.33	47.46
	RIBES	67.80	69.43	78.16	82.42
	TER	41.72	43.66	27.47	24.14
Reference S	BLEU4	27.53	23.63	41.24	30.69
	BLEU3	35.52	31.86	48.95	38.04
	BLEU+1	27.56	23.67	41.27	30.73
	RIBES	73.61	73.63	80.18	70.59
	TER	36.51	39.52	27.60	31.51
Avg. Correct Answer Rates		0.524	0.696	0.693	0.875

3.3 各問の翻訳品質と正答率

内的評価の結果と外的評価の結果（正答率）の関係性について分析した．自動評価尺度に加え，翻訳の質についての人手評価を行った．正答率，自動評価，人手評価，全ての尺度を順序尺度と捉えた上で, 同一問題に対する訳を G-Y，Y-S，S-O 間で比較した際の，評価値の大小関係の一致率を測定した．

(1) 翻訳の質についての人手評価

5 名の日本語母語話者（以下，人手評価者）が調査資料 40 問各問について，4 種のシステムの訳に同位を許して順位を付けた．人手評価者は，一度につき 1 つの英語問題についての 4 種の和訳を提示され, 翻訳の質の良さについて, ”G < Y < S = O”のように，相対的な順位を付ける．この評価手法は，the Joint 5th Workshop on Statistical Machine Translation and Metrics for Machine Translation における手法[2]に倣ったものである．5 名の人手評価者による相対順位は，全て 6 つ(=4C2)の二項関係に分割される．各二項関係の中で他の

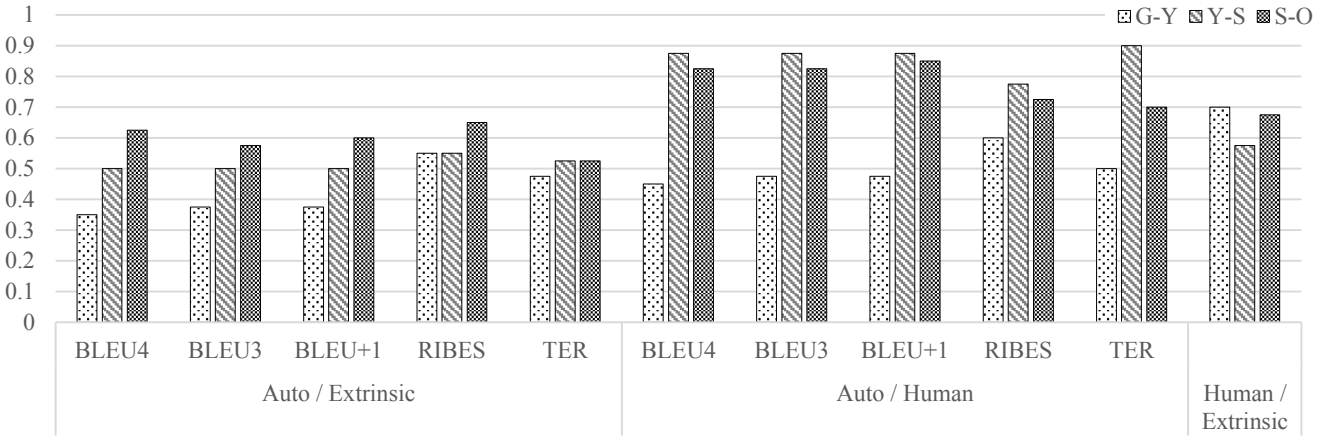

図 4: 3種の評価尺度の間で各問に対する評価値の大小関係が一致する割合

システムより優位に立つシステムがあった場合に優位のシステムに 1 点が与えられる．各システムの総点を問題別に集計し，最終的なシステム間順位を決定する．

(2) 評価尺度間の整合性

評価尺度の間で 40 問各問に対する評価値の大小関係が一致する割合を，3 種の評価尺度のペアワイズセット毎に測定した．図 4 に，左から「自動評価値と正答率」,「自動評価値と人手評価値」「人手評価値と正答率」の間の大小関係の一致率を示す．

図 4 から，自動評価値と正答率の大小関係の一致率は，全て 0.65 を下回ることがわかる．両尺度について無作為に評価した場合の大小関係の一致率が 0.5 であることから，自動評価尺度が対話の翻訳の解釈性を予測する良い評価尺度であるとはいい難いことがわかる．

BLEU 系評価値と人手評価値の間では Y-S, S-O 間について，TER 値と人手評価値の間では，Y-S 間について，それぞれよく大小関係が整合するといえる．しかしながら，G-Y 間で人手評価値と大小関係が十分に整合する自動評価値はない．これら自動評価値と人手評価値の整合性から，自動評価尺度の信頼性は翻訳手法の特性と評価される訳の質に大きく影響を受けることがわかる．

正答率は，自動評価値と比較して，人手評価値によりよく整合する傾向にあることが分かる．しかしながら，人手評価値と正答率の大小関係の一致率はすべてのシステムペアにおいて 0.7 以下である．このことは,「翻訳の質の評価」と「対話の理解の度合いを反映した評価」の間には不一致が存在し，ある特定の翻訳誤りによって対話の理解が決定的に妨害される可能性があることを示唆する．この点について，次節で検討する．

3.4 翻訳誤りの分析

翻訳誤りがどのように正答率に影響するかに基づき，翻訳文に”g”と”e”の 2 種のタグを付した．”g”は，非文法的であることを示す．”e”は文法的であるが，解釈できないあるいは誤った解釈を招く可能性のある訳であることを示す．

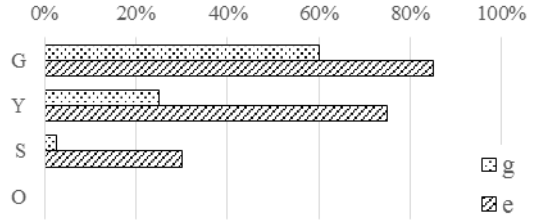

図 5: 一つ以上の誤りタグが付された文を含む翻訳済問題の割合（翻訳システム別）

表 2: Set G, Y, S 中の 120 問を「問題主要部に翻訳誤りを含むこと」及び「正答率の高低」に基づいて分類した際の各分類の問題数

Correct Ans. Rate	g		e	
	tagged	not	tagged	not
> 0.8	11	33	19	25
0.8 ≥	24	52	57	19

図 5 に，少なくとも一つの誤りタグが付された文を含む翻訳済問題の割合をシステム別に示す．

Set G, Y, S 中の 120 問について,「会話部もしくは正答選択肢（以下，問題主要部）に誤りタグ(g or e)を付された文が含まれるか否か」,「正答率が 0.8 以下であるか否か」を基準に分類した．表 2 に，各分類の問題数を示す．表 2 について，フィッシャーの正確確率検定を行ったところ，問題主要部に”e”が付された文が含まれることと正答率が 0.8 以下であることの間に，有意な交互作用がみられた($p < 0.05$)．一方，”g”を含む文があることには，有意な交互作用がみられなかった($p = 0.534$)．このことから，対話理解においては,「非文法的な文」に比べて「文法的であるが誤りを含む文」の方が，より誤った解釈を導きやすいことがわかる．

4. 関連研究

機械翻訳の補助を受けた対話や文字ベースチャットを通じて，外的評価を行う先行研究がいくつかある[5][12][14]．これらの研究は，質問紙や被験者インタビューに焦点を当てた主観的な評価に基づいた分析を行ったものである．これらのように本格的に機械翻訳の補助を受けた会話システムを伴った複雑な実験プロトコルは，本研究のような単純なプロトコルと比較して，客観的な評価に資するサンプルを多く収集することを困難にする．

他のアプローチとして，機械翻訳出力を事後編集する際のコストに基づいて外的評価を行う研究が挙げられる[4][8]．“translator’s amanuensis”としての機械翻訳の実用性を評価するタスク設定といえる．これらの研究は，第二言語に一定の能力を持つ者（翻訳の専門家やポストエディタ）が機械翻訳出力を応用するケースを想定しており，mono-lingual なエンドユーザが機械翻訳の補助を受ける際のパフォーマンスの測定を趣旨とする本研究とは異なった結論が導き出されることが推測される．

言語横断情報検索タスクもまた機械翻訳の典型的な外的評価タスクであり，非常に実用的な機械翻訳の応用方法といえる[3][11]．しかしながら，検索対象文書ないし検索クエリを適切に翻訳することが優先的に重要とされる当該タスクにおいて，本研究で議論対象としたような「翻訳された文書の解釈の容易さ」は必ずしも重要とは限らない．

5. おわりに

本稿では，機械もしくは人間によって翻訳された第二言語能力の試験問題に解答するタスクを通じて機械翻訳の外的評価を行う方法について述べ，評価と分析の結果を述べた．

4 種の翻訳手法を比較し，機械翻訳技術または機械翻訳システムと人間の協働様式における将来の改良に望まれるいくつかの要因を明らかにした．最も重要な要因として，「文脈を考慮した個々の文の翻訳」の重要性を示した．

今後の課題として，読解力を測る問題など他の種類の言語能力テスト問題での調査，機械翻訳に補助されたコミュニケーションの実現に向けた，よりインタラクティブな man-machine cooperation の研究などが挙げられる．

謝辞 調査にご協力頂いた中学校の教員・生徒の皆様，調査資料として模擬試験の過去問題をご提供いただいた学校法人高宮学園代々木ゼミナールに，謹んで感謝の意を表する．

参考文献

1) Bar-Hillel, Y. 1960. Automatic Translation of Languages. In Franz Alt, Advances in Computers. Academic Press, New York.

2) Callison-Burch, C.; Koehn, P.; Monz, C.; Peterson, K; Przybocki, M.; and Zaidan, F. O. 2010. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, 17–53.

3) Fujii, A.; Utiyama, M.; Yamamoto, M.; and Utsuro, T. 2008. Overview of the Patent Translation Task at the NTCIR-7 Workshop. In *Proceedings of NTCIR-7*, 389–400.

4) Green, S.; Heer, J.; and Manning, D. C. 2013. The Efficacy of Human Post-editing for Language Translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI’13)*, 439–448.

5) Hamon, O.; Mostefa, D.; and Choukri, K. 2007. End-to-end Evaluation of a Utterance-to-utterance Translation System in TC-STAR: In *Machine translation summit XI*, 223–230.

6) Isozaki, H.; Hirao, D. K.; Sudoh, K.; and Tsukada, H. 2010. Automatic Evaluation of Translation Quality for Distant Language Pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP-2010)*, 944–952.

7) Kay, M. 1980. The proper place of men and machines in language translation, Technical Report CSL-80-11 Xerox Palo Alto Research Center, USA.

8) Koehn, P. and Germann, U. 2014. The Impact of Machine Translation on Human Post-editing. In *Proceedings of Workshop on Humans and Computer-assisted Translation*, 38–46.

9) Lin, C.-Y. and Och, F. J. 2004. Orange: A Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING2004)*, 501–507.

10) Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation: In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL’02)*.

11) Parton, K.; Habash, N.; and McKeown, K. 2012. Lost and Found in Translation: The Impact of Machine Translated Results on Translingual Information Retrieval. In *Proceedings of Association for Machine Translation in the Americas (AMTA’12)*.

12) Schneider, A.; Van der Sluis, I.; and Luz, S. 2010. Comparing Intrinsic and Extrinsic Evaluation of MT Output in a Dialogue System. In *Proceedings of the International Workshop on Spoken Language Translation*, 329–336.

13) Snover, M.; Dorr, B.; Schwaltz, R.; Micciulla, L.; and Makhoul, J. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas (AMTA’06)*.

14) Yamashita, N.; Inaba, R.; Kuzuoka, H.; and Ishida, T. 2009. Difficulties in Establishing Common Ground in Multiparty Groups using Machine Translation. In *Proceedings of CHI' 09*, 679–688.