機械翻訳における人の振る舞いを用いた誤訳検出

○向井涼　田中高志　大谷雅之（近畿大学）

Detecting Translation Error Using Human Behavior in Machine Translation

* R. Mukai, T. Tanaka and M. Otani (Kindai University)

Abstract－ In this study, we analyzed relationship between human behavior and machine translation. We have tackled the following points. ① We classified human behavior data which are obtained when subject looks at mistranslation using deep learning. ② We propose the method of detecting the mistranslation by combining the automatic evaluation method and the manual evaluation. We obtained human behavior data through experiments, (1) it is difficult to estimate the mistranslation only by the expression data, (2) the BLEU method gives a high evaluation value to the mistranslation and adds evaluation methods of different scales.

Key Words: Machine translation, Human Behavior, Deep learning

1　はじめに

近年，我が国では外国人観光客が増加の一途をたどっている．2020 年の東京オリンピックや 2015 年の大阪万博も控えており，外国人旅行者向けの多言語サポートが重要である．その一例としてデジタルサイネージを用いた多言語での情報提供を行う例もある[2].

多言語での情報提供に機械翻訳を用いる場合，翻訳精度が向上している一方で，誤訳の問題が未だに導入の障壁となっている．人手翻訳の場合，人間が文章の文脈を理解しながら訳すため，誤訳が発生しても前後の文との繋がりから本来の意味を推測できることがある．しかし，機械翻訳の場合は，文章全体の意味を理解せず，統計的に訳語を選択するため，前後の文を読んでも意味を理解できない誤訳が発生する.そのため，機械翻訳を用いる場合は,人手チェックが欠かせない．デジタルサイネージなど公共の情報提示機器では，できるだけ早く誤訳の発生を検出し，対処する必要がある．しかし，専門家による全箇所のチェックは人手の面でもコストの面でも現実的ではない．本研究では，この問題に対し，翻訳の専門家を必要としない誤訳検出方法について検討する．具体的には，まず，①機械翻訳の人手評価と自動評価手法の関係を分析し，自動評価の傾向から誤訳検出方法を検討する．また，②人手評価は時間を要するため，人間(非専門家)が誤訳を見た際の表情をデータ化し，その変化パターンによって誤訳を推定する方法について検討する．上記二点の分析に用いるデータ取得をするため，誤訳を見た際の表情データと人手評価を収集する被験者実験を行い，その結果を分析する．

2　機械翻訳の自動評価手法

2.1 機械翻訳の自動評価と人手評価

翻訳結果を評価する方法として，人間が原文と訳文を比較し評価する方法（人手評価）と，専門家が翻訳した訳文（参照訳）と，評価したい訳文（評価訳）をある尺度に基づいて比較し，自動で評価値を与える方法（自動評価）がある．前者の場合，評価の精度は高いが処理に時間を要する点，後者の場合，人手評価と比べて処理時間は短いが，評価精度が低くなる点が，それぞれ欠点として挙げられる．本研究の目的は，これら二つの利点を組み合わせた新しい評価手法を提案することである．まず，自動評価と人手評価の関係性を分析し，自動評価の結果から，人手評価の結果を予測できないか検討した．具体的には，複数の自動評価尺度と，人手翻訳の関係性について分析した．以降では，分析に利用した評価尺度である BLEU[1]と RIBES[3]について説明する．

2.2 BLEU

BLEUは機械翻訳の分野において，最も一般的な自動評価尺度であり,n-gramマッチ率に基づく手法である．BLEUではn-gramをn=4で用いる．1-gramは，単語訳の正しさを表す尺度を意味し，高次のn-gramは，翻訳の流暢さを表す尺を表している．翻訳文が参照訳より長い場合はペナルティがなくなり，影響を及ぼさない．逆に翻訳文が参照訳より短い場合，その短さに応じてペナルティが大きくなり，評価値が小さくなる．BLEUはこの両者を組み合わせた尺度である．BLEUの評価値は0.0から1.0の実数で表現され，この値が高いほど参照訳に近い翻訳文と見なされる．

BLEUの課題として,4単語までの連続した短い単語列しか評価ができないため，翻訳文と参照訳の意味が大きく異なっていても局所的に参照訳の単語列が存在すると評価値が高くなるという問題がある．逆に，意味が合っていても語順の違いにより，評価が低くなることがある．

2.3 RIBES

BLEU における語順の誤りに対して正しい評価をすることができないという弱点に対し，RIBES と呼ばれる単語の並びを重視する評価尺度が提案されている．日英・英日翻訳では並び替えが発生する言語対において，文を正しい並びで表現するのは，文の意味を正しく伝えるためには重要である．RIBES は正規化された順位相関係数を用いて評価する．順位相関係数は評価訳と参照訳の間で共通単語にのみ着目して計算する．そのため，2 単語のみなど単語数が極端に少ない場合で語順の入れ替えもない場合，内容が異なる訳でも，順位相係数が 1 となり評価が高くなる問題が発生する．この問題を，翻訳文と参照訳に含まれる単語を含む割合をペナルティとして導入することで回避している．

一般的には原文に対する正解訳は複数あることが多く，BLEU も複数の参照訳を前提に設計されている指標であるため，RIBES でもこれに対応することが課題となっている．

3　自動評価と人手評価の関係分析

ここでは，高速かつ精度の高い自動評価尺度の実現を目指し,複数の自動評価と人手評価の関係を分析し，誤訳の検出方法を検討した結果について述べる．

3.1 データの取得

まず，分析に使うデータとして，英日翻訳を対象として人手評価と自動評価の収集を行った．原文となる英文と参照訳となる和文は，英語の問題集[4]から引用した．評価訳となる翻訳文は，Google 翻訳と，京都言語グリッドに登録されている J-Sever[5]をそれぞれ用いて生成した．生成された翻訳文の評価値を BLEU と RIBES を用いて算出した．以降それぞれの評価値を BLEU 値，RIBES 値とする．

上記のデータを収集するために被験者実験を行った．具体的には，各被験者に PC のディスプレイ上に評価訳と参照訳を 15 秒間表示したのち，5 秒間で評価訳を三段階評価する,ということを各英文について行った．被験者は，近畿大学理工学部に所属する学部生 4 名(被験者 A〜D)である（TOEIC の平均点：358）．実験に使う英文の数は Google 翻訳を用いた実験を 250 文，同様に J-Server を用いた実験を 250 文の合計 500 文である．以下に 3 段階評価の詳細を示す．

2：翻訳文のみで完璧に理解できた．
1：翻訳文と参照訳を比べたら理解できた
0：参照訳を見ても翻訳文を理解できなかった

3.2 結果と考察

この章では，実験結果とその考察を述べる．被験者 B を一例として，Google 翻訳の翻訳文と参照訳を比較した実験結果を Fig. 1 に，J-Server の翻訳文と参照訳を比較した実験を行ったときの実験結果を Fig. 2 に示す．青，橙，灰色の各点は，人手評価がそれぞれ 0，1，2 であったときの英文について BLEU 値を横軸， RIBES 値を縦軸としてプロットした結果を示している．

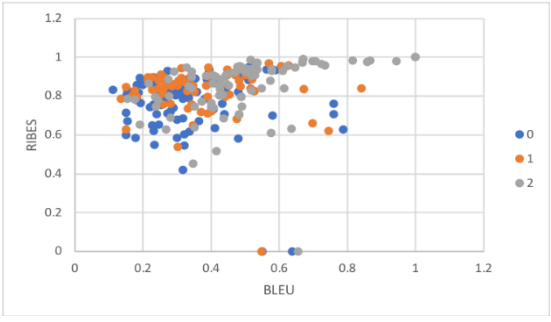

Fig. 2: BLEU v.s. RIBES (Google translation)

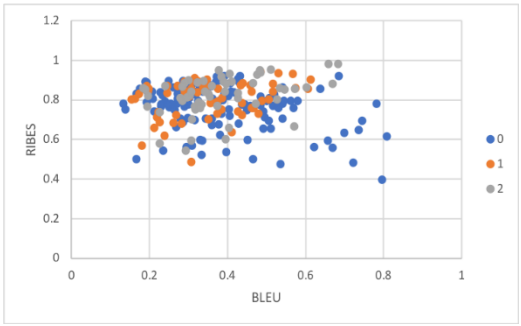

Fig. 1: BLEU v.s. RIBES (J-Server)

Fig. 1 を見ると，Google 翻訳の場合，右肩上がりのグラフになっており，BLEU と RIBES の評価に相関がみられる．人手評価についても，BLEU 値と RIBES 値が高いものは高い評価値となっており相関が見られるものの，RIBES 値はどの評価訳についても 0.4 以上となっており，人手評価が低いものについても高い評価値が与えられており，分類を困難にしている．Fig. 2 の J-Server の場合,RIBES 値と人手評価に相関がみられる．しかし，BLEU 値は，人手評価と関係なく分布されており，誤訳にも高い評価を与えている．すなわち，J-Server を用いて生成された翻訳文については,BLEU を用いて人手評価を推測することは困難であると言える．これらの結果は，機械翻訳サービス毎に人手評価と相関する自動評価が異なり，翻訳サービスに適した自動評価手法の選択が重要であることを示唆している．

4　人の振る舞いの分析

BLEU や RIBES などの自動評価手法を用いた誤訳推定はテキストベースで行うアプローチであるため精度に限界がある．そこで本研究では，誤訳を見た際の人の振る舞いから誤訳を検出する方法も検討した．具体的には，人が機械翻訳された訳文を見た際の，誤訳が発生した場合の表情と，発生していない場合の表情を分析し，誤訳の検出方法を検討した．また，深層学習を用いた分類も試みた．

4.1 人の表情のデータ化

人の表情を数値化するため，本研究では Microsoft 社のクラウドプラットフォーム Azure で提供されている Face API を利用した．Face API は，画像に含まれている人の顔を検出，認識，分析するためのサービスであり，人の顔が写っている画像を送信すると，JSON 形式のデータで画像の分析結果を返す．Face API の機能としては，送信した画像に写っている人の性別，年齢，メガネ着用の有無，表情などの顔に関連する属性を抽出できる．本研究では，写真に写っている人物の表情のみ使用した．Face API が返す表情値は 7 種類に分かれており，sadness（悲しみ），neutral（中立），contempt（侮辱），disgust（嫌悪），surprise（驚き），fear（恐れ），happiness（喜び）のそれぞれの値を，合計 1.0 となる 0 以上の実数で返す．

4.2 データの取得

機械翻訳された訳文（評価訳）の表情データを取得するために被験者実験を行った．3.1 と同様に，問題集から引用した英文とその参照訳，および評価訳をそれぞれ 50 文用意した．被験者には参照訳と機械翻訳された訳文を順番に 15 秒間読ませ，その間 1 秒間に 1 枚被験者の顔が正面から写るように写真を撮影する．被験者は，その後 5 秒間で，読んだ参照訳と評価訳を比べて誤訳であるかどうかを判断するということを繰り返した．被験者は 3.1 同様，近畿大学理工学部の 4 名の学生を対象とした．

4.3 結果と考察

被験者実験で取得した顔写真から Face API を使用して表情データを取得した．ここでは，被験者 A の表情データを一例として，10 文の参照訳と評価訳を見せた結果とその考察について述べる．

Fig. 3 は誤訳と感じなかった場合の表情の時系列データ 10 文分のデータをグラフにプロットしたものである．横軸は訳文を表示してからの経過時間（秒）を，縦軸がその際の表情データ値を表している．四角の点が sadness の値を，丸点が neutral の値を表している．点の色の違いは文の違いを意味している．Fig. 4 も Fig. 3 と同様だが，被験者 A が誤訳と感じた場合のグラフを表している．これらの図から，誤訳と感じなかった場合の表情データに関しては表情値の変化が少ない．一方で，誤訳と感じた場合の表情データでは neutral の数値が一度下がってから元に戻る(sadness は上がって戻る）という変化を生じている．これは，被験者 A の場合，誤訳を見た際に表情が大きく動くことを意味している．表情値が大きく動くということは，15 秒間の表情値の分散も大きくなることが予想される．そこで，各被験者の neutral の平均と分散を比較した．

Table 1 に，被験者 4 人の誤訳と感じた場合とそうでなかった場合の neutral の平均と分散を示す．上から順に，誤訳と感じなかった場合の平均および分散，誤訳と感じた場合の平均と分散を，各被験者について記載している．この表から，被験者 D 以外の三名については，誤訳と感じた場合の分散値が高く，誤訳と感じない場合の分散値が低いという傾向が表れた．

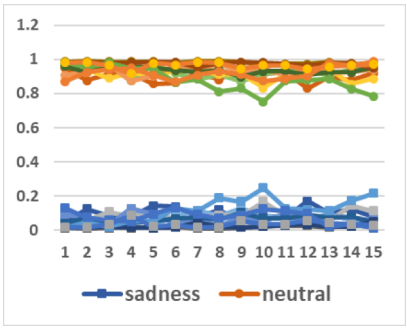

Fig. 5 : Face expression Graph (correct translation)

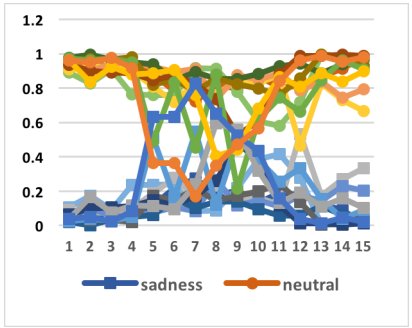

Fig. 4 : Face expression Graph (mistranslation)

	subject A (neutral)	subject B (neutral)	subject C (neutral)	subject D (neutral)
Average (correct trans.)	0.9181200	0.8724928	0.8242396	0.7530275
Variance (correct trans.)	**0.0005806**	**0.0909227**	**0.0001309**	0.2092168
Average (mis-transla-tion)	0.8237733	0.8621210	0.8890308	0.6763152
Variance (mis-transla-tion)	**0.4605014**	**0.5524244**	**0.6246938**	0.0100579

Table 1 : Averages & Variances corresponding to the neutral values of subjects

4.4 深層学習による分類

前章の考察から，被験者の表情データの neutral には誤訳を見た場合とそうでない場合で傾向が異なる場合があった．そこで、被験者 A の表情データを使用して深層学習を行い，実際に分類できるか確認した．
具体的には，深層学習の中でも時系列データの分析に特化した Recurrent Neural Network を使用した．被験者 A の表情データを 40 個(40 文)文のデータを入力データとして，各訳文に対しての誤訳と感じたかそうでないかの評価データをラベルデータとしてモデルを学習させた．その後，学習に使用していない被験者 A の表情データ(Fig. 5 は人手の評価値 0，Fig. 6 は人手の評価値 1)を，学習したモデルを用いて分類を試みた．Table 2 はこの分類結果を表している．

Table 2 より Fig. 5 に対しては被験者が誤訳と感じたので評価値が 0 になっているのに対し評価値の予測は 0.0006438 とほぼ 0 に近く，予測ができている．

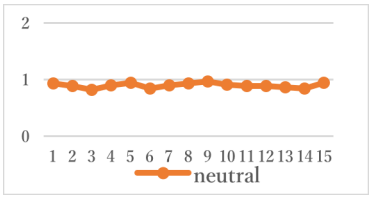

Fig. 5 : Face expression test (correct translation)

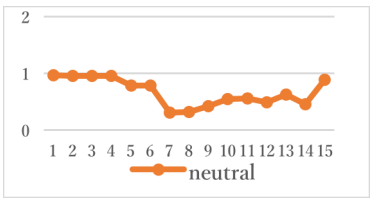

Fig. 6 : Face expression test (mistranslation)

Fig.6 に関しては被験者の評価値が 1 に対し評価値の予測は 0.8121569 と予測値の約 8 割となった.

以上より，訳文を見て誤訳と感じた場合と感じなかった場合の人の表情データを用いて深層学習を行うことで，誤訳を判定することができる場合があることが明らかになった．この結果は，訳文を見た際の人の表情値にパターンがある場合だけであるため，被験者 D のように，表情値に変化がない人の場合の誤訳推定方法が今後必要である.

Table 2 : Translation Score Prediction

	Fig 5	Fig 6
Label(Evaluation score)	0(mistranslation)	1(correct trans)
Predicted score	0.0006439	0.8121569

5　まとめ

本研究では，高速かつ精度の高い自動評価尺度を目指し，以下の 2 点を試みた.

①複数の自動評価と人手評価を組み合わせた誤訳検出方法の検討

②人の振る舞いデータを用いた誤訳判定方法の検討

あ①について，BLEU と RIBES の自動評価手法の組み合わせによる誤訳検出方法を検討した結果，人手評価は，RIBES とは相関があり，誤訳検出に適性があるが，BLEU は，人手評価と相関がなく，誤訳に高い評価を与えることがあることが明らかになった． また，RIBES は，人手評価と相関があったが，RIBES 値の分布が偏ってより，詳細な分類は困難であった．これにより，機械学習を用いて誤訳検出を行う場合，BLEU と異なる尺度の評価手法を加える必要があることがわかった．さらに，機械翻訳サービス毎に適した自動評価手法があることも明らかとなった.

②について，BLEU や RIBES に変わる誤訳検出の手法を提案した．機械翻訳された訳を見た時の人の表情を使用して，その訳について人手評価し，それらに対し機械学習を用いることによって，人の表情だけで誤訳検出を試みた．被験者 1 名のデータで学習させ，テストしたところ，誤訳検出の予測は可能だったが，表情にパターンが現れない被験者の場合は分類が困難であった．また，人によって表情のでるカテゴリ，数値の大きさ,法則性が異なることがあることが分かった.

今後の課題としては，①について，RIBES と BLEU の他の自動評価手法との関係についても人手評価との関係性を分析する必要がある．また，機械翻訳機に適した自動評価手法を選択するための方法を検討する必要がある．②については，表情データを用いての翻訳品質の評価をおこなったが，複数人数の表情データを使用し学習させ分類を試みる必要がある．また人によって表情にパターンが現れないことがあるが，人の振る舞いは表情だけではなく，視線なども考えられる．そのため，訳文を読んでいる時の人の視線などをトラッキングすることにより，人が誤訳を見ている時のパターンが見つかる可能性があると考えられる．表情以外の人の振る舞いを使った分類も考えることが今後の課題になっていく.

6　謝辞

本研究は，日本学術振興会科学研究費基盤研究(B)(18H03341,平成 30 年度～32 年度)の助成を受けた.

7　参考文献

1) Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. 40th Annual meeting of the Association for Computational Linguistics,pp. 311-318, 2002.
2) 小荷田樹之，小木哲朗，& 廣井裕一．(2016)．多言語表示に自動対応するデジタルサイネージシステムの開発．第 78 回全国大会講演論文集，2016(1)，31-32.
3) 平尾努・磯崎秀樹・Kevin Duh・須藤克仁・塚田元・永田昌明．RIBES:順位相関に基づく翻訳の自動評価 法．言語処理学会 第 17 回年次大会 発表論文集，pp. 1115-1118，2011.
4) 篠田重晃，瓜生豊．(2011)．Next Stage 英文法・語法問題 3rd edition．桐原書店.
5) Toru Ishida Ed. The Language Grid: Service-Oriented Collective Intelligence for Language Resource Interoperability. Springer, 2011. ISBN 978-3-642-21177-5.