英日機械翻訳における既存の評価指標の線形結合による性能評価

中屋和樹[1] 川又泰介[1] 松田源立[1]
[1]成蹊大学 理工学部
us182092@cc.seikei.ac.jp {kawamata,matsuda}@st.seikei.ac.jp

概要

機械翻訳の自動性能評価指標として様々な手法が提案されているが、特に構文が大きく異なる英日間の翻訳では、どの手法が優れているかについて明確に評価することは困難であった。本研究では同一の英文に対する Google 翻訳と DeepL の出力和文を実験用データとし、DeepL の結果の方が翻訳性能が高いと仮定した。そして、その 2 出力文を正確に判別できる指標を優れた指標であるとみなした。実験では、BLEU 等の n-gram ベースの指標、MER 等の編集距離ベースの指標、BERTScore 等の分散表現ベースの指標等を用いて、その判別の正解率を比較した。さらに、SVM によりそれらの評価指標を線形結合した場合の正解率も比較した。その結果、異なる系統の評価指標を線形に組み合わせることで、判別の正解率を向上させることができることを示した。

1 はじめに

近年、深層学習の発展に伴いニューラルネットワークを用いた機械翻訳モデルの研究が盛んに行われている[1]。これら翻訳モデルの精度を測るには、人手評価との相関が高い自動評価指標が必要となる。現在、自動評価指標のディファクトスタンダードは BLEU スコア[1]であり、機械翻訳の論文において提案システムと既存システムの比較に用いられることが多い。しかし、BLEU は弱点が多く指摘されている。例えば、参照文の局所的な n-gram をモデル出力文が保持していた場合、意味が通っていなくても高い評価を与えてしまう[2]。構文が似ている言語間では翻訳モデルがこのような出力をすることは少ないが、日英翻訳のような構文が大きく異なる言語間では頻繁に見られる。また、意味的に近い語彙の言い換えが発生した際に対応することができないので、低い評価になってしまうこともある[3]。語彙の言い換えはどの言語でも起こりうることであり、これに対応できなければ翻訳モデルに対する正しい評価ができない。評価指標の優劣を決めることは非常に難しく、BLEU が発表されて以降、様々な自動評価指標が提案されており、上記の BLEU の弱点に対応できるものもある。しかし、それらの指標はそれぞれが長所と短所を持っており、最適な指標というのは定まっていない。

現在翻訳ツールで有名なものは「Google 翻訳[4]」と「DeepL[5]」であるが、DeepL の方が Google 翻訳に比べて、読みやすく、理解しやすい文章を出力するとされている[6]。そこで、本研究では、ある評価指標をもとに Google 翻訳と DeepL を判別できれば、その評価指標は良いものであると仮定した。そして、同一英文群に対する Google 翻訳と DeepL の出力和文群について様々な翻訳評価指標の値を算出し、判別の正解率を比較した。次に、それらを特徴量として線形 SVM で最適な組み合わせを学習し、どのような組み合わせが良い指標となるかを調査した。さらに、それらの分析結果と具体的な出力文と照らし合わせることで、それぞれの評価指標の長所・短所の考察を行う。

2 先行研究

本研究では、様々な観点から見た評価尺度を利用する。本節ではそれらについての簡単な説明を行う。今回用いる評価指標の全体的な分類を表 1 に示す。

表 1 本研究で用いる評価指標

n-gram	BLEU, ROUGE1, ROUGE2, RIBES
編集距離	WER, MER, CER
分散表現	WMD, WRD, BERTScore

2.1 n-gram ベース

・BLEU BLEU は現在、機械翻訳の自動評価指標として最も用いられている。2002 年にオリジナルの論文が発表されて以来、改良された BLEU がいくつか出ているが、今回は BLEU+1[7]を用いる。基本的な

考えとしては、モデル出力文と参照文の n-gram の一致度によって翻訳の精度を評価する指標である。

・ROUGE　ROUGE-n は 2003 年に Chin ら[8]によって発表された指標であり、元々は要約タスク専用の評価指標である。具体的には、参照文とモデル出力文について Recall と Precision を計算し、最後に F1 値を計算することによって、ROUGE スコアとしている。ROUGEn の n は使用する n-gram を表しており、本研究では ROUGE1 と ROUGE2 を用いた。

・RIBES　RIBES は 2011 年に平尾ら[2]によって発表された翻訳評価指標であり、順位相関係数を利用した評価指標である。

2.2 編集距離ベース

・WER　WER(Word Error Rate)は通常は音声認識の評価指標として用いられているものであり、入力文に対する誤ったトークン数の割合で表される

・MER　MER(Match Error Rate)は WER の問題点を解決するために、2004 年に Andrew ら[9]によって提案された。具体的には、WER の式の分母に挿入数を追加する。

・CER　CER(Character Error Rate)は基本的な考え方は WER と同じで、編集距離に基づいたエラー率を計算する。WER はトークン単位での変換を行うが、CER では文字単位(Character)での変換を行う。

2.3 分散表現ベース

・WMD　WMD(Word Mover's Distance)[10]は、分散表現を用いて最適輸送問題を解くことにより、2 つの文書の類似度を測る評価指標である。

・WRD　WRD (Word Rotator's Distance)は 2020 年に横井ら[11]によって発表された WMD の改良版であり、コサイン類似度を利用している。

・BERTscore　BERTscore は 2020 年に Tianyi ら[12]によって提案された、2 つの文書間の類似度をはかる指標である。具体的には、2 文の分散表現ベクトル集合により、Precision、Recall、F 値を計算する。本研究では F 値を採用した。

3 実験設定

・翻訳データ　評価指標を計算するためには、翻訳した文に対する正解データ(リファレンス)が必要となる。よって、対訳コーパスである ASPEC コーパス[13]の文章データを用いた。学習データ・開発データ・開発試験データ・試験データの 4 種類があるが、今回は試験データ(1812 文)全てを選択した。

・翻訳ツール　先述の通り「Google 翻訳」「DeepL」を利用した。前者は googletrans[14]、後者は手動で文章を入力して翻訳した。

・分散表現　一部の評価指標を計算する際には分散表現が必要となる。今回は、ボキャブラリ数を 32000 に設定し、JParaCrawl コーパス[15]を用いて事前学習を行った Transformer[16]のデコーダ層の分散表現を用いた。モデル自体は fairseq[17]で公開されているものである。

・トークナイザ　トークナイザは文章をトークンに分割する処理を行うものである。本研究では Sentencepiece[18]と MeCab[19]を用い、各々について評価指標を計算した。ただし、CER については文字単位のためトークナイザは不要であり、分散表現ベースの指標については Sentencepiece のみを利用した。

・評価指標　評価指標としては、表 1 に示した 10 種類について Sentencepiece を適用したものを採用した。さらに、n-gram ベースの全てと編集距離ベースの WER と MER については MeCab を適用したものも採用し、計 16 種類となった。

・線形学習モデル　線形 SVM を利用した。具体的には、文レベルでそれぞれの評価指標の値を算出し、それらを特徴量として学習を実行し、Google 翻訳と DeepL の線形 SVM による 2 値判別を行った。データは、学習データ 90%(1631 文)とテストデータ 10%(181 文)に分割した。また、学習の際に Optuna[20]を用いることにより、学習データのみでハイパーパラメータをチューニングした。

4 結果と考察

4.1 評価指標間の相関

図 1 は、Google 翻訳と DeepL の 3624 件の全出力文上で各評価指標間の相関係数行列を計算し、ヒートマップで表したものである。相関係数は最低でも 0.5 以上であり、互いにある程度の相関を持つことは確かめられた。それ以外にも以下のような関係性が観察できた。第一に、RIBES_Sent と RIBES_MeCab は n-gram ベースでありながら、BLEU や ROUGE との相関が低く、代わりに編集距離系の WER や MER と相関が高い。これは RIBES がユニグラムの一致を図りつつも、その順番を重視した評価指標であるからだと考えられる。編集距離系の評価指標は先頭から順番に置換・削除・挿入を行なっていくため、語

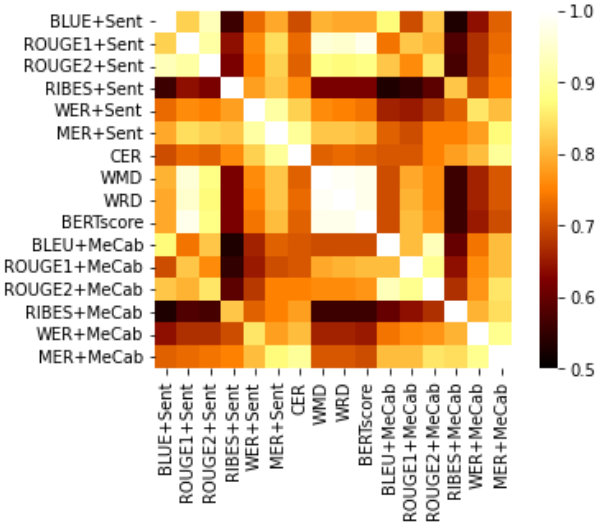

図 1　各評価指標の相関係数行列

表 2　単体での判別正解率

	acc(%)
BLEU_Sent	64.19
ROUGE1_Sent	66.12
ROUGE2_Sent	65.01
RIBES_Sent	62.53
WER_Sent	**67.49**
MER_Sent	66.77
CER	64.74
WMD	66.67
WRD	65.56
BERTscore	66.39
BLEU_MeCab	63.91
ROUGE1_MeCab	63.91
ROUGE2_MeCab	62.26
RIBES_MeCab	64.19
MER_MeCab	66.94
WER_MeCab	64.74

順に対して厳しい評価をする。Google 翻訳は英語を直訳したような文章になることが多く、人手で翻訳したものと比べると、内容が反対になっていることがある。よってこのような文章に対して、編集距離系は高いエラー率(低い評価)を出力し、RIBES では順位相関係数が負になった(低い評価)ため、比較的相関が高くなったのだと考えられる。第二に BLEU_Sent と BLEU_MeCab は相関が高い評価指標に違いがあり、前者は n-gram+Sent、分散表現、編集距離+Sent、後者は n-gram+MeCab、編集距離+MeCab

表 3　2 つ組み合わせた場合の判別正解率

	acc(%)
ROUGE1_Sent + RIBES_MeCab	**68.87**
BLEU_Sent + MER_Sent	68.60
RIBES_MeCab + BERTscore	68.32
MER_MeCab + WMD	68.04
ROUGE1_Sent + MER_Sent	68.04

と相関が高くなっている。分散表現ベースの評価指標は全て Sentencepiece でトークナイズしていることから、同じトークナイザでトークナイズして算出した評価指標は相関が比較的高くなると考えられる。

4.2 判別正解率

表 2 は、特徴量を 1 個のみ(各指標単体)とし、線形 SVM で閾値のみを学習した時の判別正解率である。WER_Sent が最も高い正解率であった。表 3 に 2 つの評価指標を特徴量として組み合わせの重みを学習した場合の正解率を上位 5 位まで示す。表 4 に 3 個もしくは 4 個の評価指標を特徴量とした場合の正解率を上位 10 位まで示す。この結果から、評価指標単体より複数の評価指標を組み合わせた方が正解率が向上することが分かる。表 4 における各評価指標の登場回数を見ると、BERTscore が 7 回、ROUGE1_Sent が 6 回、RIBES_MeCab が 6 回で上位となる。これらの評価指標は表 3 にも頻繁に登場している。図 1 の結果とも併せて考えると、分散表現ベースである BERTscore、単純な n-gram ベースである ROUGE1、n-gram ベースではあるが編集距離ベースと性質が近い RIBES の組み合わせがより有効な評価指標となることが示された。以上が統計的な結果であるが、具体的な文例での組み合わせの有効性については、付録 A にて紹介している。

4.3 まとめ

本研究によって得られた結果も踏まえた各評価指標の特徴(長所と短所)を表 5 に示す。現在最も使用されている BLEU は、直感的に理解できプログラミングもしやすいが、クリティカルな短所が多いため、率先して使用すべきではないと考えられる。もし n-gram 系を使用するのであれば、ユニグラムのみでもある程度評価はできるため、ROUGE1 や RIBES が良いと考えられる。編集距離系の評価指標は、表 2 より WER_Sent の正解率が一番高いこと、表 4 より

表 4　3、4 つ組み合わせた場合の精度

	acc(%)
ROUGE1_Sent + RIBES_MeCab + BERTscore	**69.15**
ROUGE1_Sent + RIBES_Sent + MER_Sent	68.87
BLEU_Sent + RIBES_MeCab + MER_Sent + BERTscore	68.87
ROUGE1_Sent + RIBES_MeCab + WER_Sent +WRD	68.87
ROUGE1_Sent + RIBES_Sent + MER_Sent +BERTscore	68.87
ROUGE1_Sent + RIBES_MeCab + MER_Sent +BERTscore	68.87
ROUGE1_Sent +RIBES_MeCab + WER_Sent +BERTscore	68.87
BLEU_Sent + RIBES_Sent + MER_Sent	68.60
BLEU_Sent + RIBES_MeCab + BERTscore	68.60
RIBES_Sent + MER_Sent +BERTscore	68.60

表 5　各評価指標の特徴

	長所	短所
n-gram	・BLEU と ROUGE は直感的にわかりやすい ・単純なユニグラムの ROUGE1 や RIBES でもある程度は妥当な評価ができる ・評価値の範囲が 0~100%である	・BLEU は一般に評価値が低く、段階的な評価が困難である ・逆に RIBES は評価値が過剰傾向にある ・日本語のように分割方法が曖昧な言語では、高次 n-gram が一致しない ・語彙の言い換えに対応できない
編集距離	・語順に対して厳密な評価ができる ・MER、CER は範囲が 0~100%である	・意味的に同一でも語順の異なる文を正しく評価できない ・語彙の言い換えに対応できない
分散表現	・語彙の言い換えに対応可能である ・表面上の単語の一致にとらわれない評価が可能である ・単語アラインメントが利用できる	・評価値は、使用した分散表現に依存する ・WMD、WRD におけるアラインメント計算は高コストである ・評価最低値、最高値はデータに依存する

WER_Sent と MER_Sent が上位に現れていることから、WER と MER は翻訳評価指標として用いることが可能であると考えられる。ただし、表 5 より、語順の維持に強く依存することに注意する必要がある。また結果から RIBES で代用可能である可能性があるが、さらなる分析が必要である。

分散表現系の評価指標は、分散表現に大きく依存するという短所があるが、語彙の言い換えに対応可能であり、また、単語アラインメントを得られるという大きなメリットがある。特に BERTscore はアラインメントが WMD や WRD に比べて容易に計算可能であり、4.2 節において良い評価指標であるということが判明している。よって、分散表現系は BERTscore を用いるのが良いと考えられる。

5 終わりに

本研究では、英日翻訳のコーパスに置いて様々な翻訳性能の評価指標を調査し、それらを特徴量として Google 翻訳と DeepL の線形 SVM による 2 値判別を行った。その結果、異なる種類の評価指標を線形に組み合わせることで判別正解率が向上することが判明した。さらに各評価指標の特徴について考察した結果、n-gram ベースでは ROUGE1、RIBES、編集距離ベースでは WER、MER、分散表現ベースでは BERTscore を用いるのが良い、という結論が得られた。今後はデータセットの拡張、指標の追加と実験を通して、より適切な評価指標の構築を目指す予定である。

謝辞

本研究は JSPS 科研費 JP21K12036 の助成を受けたものである。

参考文献

1. Kishore Papineni et al. “BLEU: a Method for Automatic Evaluation of Machine Translation”. In: ACL, Philadelphia, July, 2002, pp. 311-318
2. 平尾努 他 “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 (2011 年 3 月)
3.http://cr.fvcrc.i.nagoya-u.ac.jp/~sasano/pdf/snlp2019sasano.pdf (2022 年 1 月 11 日 アクセス)
4. https://translate.google.co.jp/?hl=ja (2022 年 1 月 11 日 アクセス)
5. https://www.deepl.com/ja/translator (2022 年 1 月 11 日 アクセス)
6. Ahmad Yulianto “Google Translate vs. DeepL: A quantitative evaluation of close-language pair translation” In: AJELP, ISSN 2289-8689 / e-ISSN 2289-8697, Vol 9 No.2 (2021), 109-127
7. Graham Neubig “文レベルの機械翻訳評価尺度に関する調査” Information Processing Society of Japan 2013
8. Chin-Yew Lin and Eduard Hovy “Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics ” In: Processing of HLT-NAACL, 2003, pp. 71-78
9. Andrew C. Morris el at. “From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition ” In: INTERSPEECH 2004-ICSLP
10. Matt J.Kusner et al. “From Word Embeddings To Document Distances ” In: Processing of the 32th International Conference on Machine Learning, Lile, France, pp. 957-966 2015
11. 横井祥 他 “単語埋め込みのノルムと方向ベクトルを区別した文書最適輸送コスト” The 34th Annual Conference of the Japanese Society for Artificial Intelligence, 2020
12. Tianyi Zhang et al. “BERTSCORE: EVALUATING TEXT GENERATION WITH BERT ” In: ICLR 2020
13. https://form.jst.go.jp/enquetes/aspec_order_sampleJ-T (2021 年 6 月 24 日 受付)
14. https://pypi.org/project/googletrans/ (2020 年 1 月 11 日 アクセス)
15. 森下睦 他 “JParaCrawl: 大規模 Web ベース日英対訳コーパス” 言語処理学会 第 26 回年次大会 発表論文集 (2020 年 3 月)
16. Ashish Vaswani et al. “Attention is all you need.” In: CoRR, Vol. abs/1706.03762, 2017.
17. https://github.com/pytorch/fairseq (2022 年 1 月 11 日 アクセス)
18. https://github.com/google/sentencepiece (2022 年 1 月 11 日 アクセス)
19.https://github.com/neologd/mecab-ipadic-neologd (2022 年 1 月 11 日 アクセス)
20. https://github.com/optuna/optuna (2022 年 1 月 11 日 アクセス)

付録 A. 具体的な文例における評価指標の組み合わせの効果

表 A.1　RIBES、BERTscore、MER を加えることで判別可能となった文の例(指標の単位は全て%)

	本文	BLEU	RIBES	BERTscore	MER
リファレンス(正解文)	エアータービンハンドピース(A)と,治療用及び技工用マイクロモータハンドピースについて解説し、さらに、A における新技術の展開について検討した。				
Google 翻訳	本稿では、治療用および歯科用技術用の空気タービンハンドピース(A)とマイクロモータハンドピースについて説明します。	13.76	52.12	70.77	85.29
DeepL 翻訳	本稿では、治療・歯科技工用のエアタービンハンドピース(A)とマイクロモーターハンドピースについて述べ、A の新技術の開発についても述べる。	13.99	**71.75**	**75.47**	**73.38**

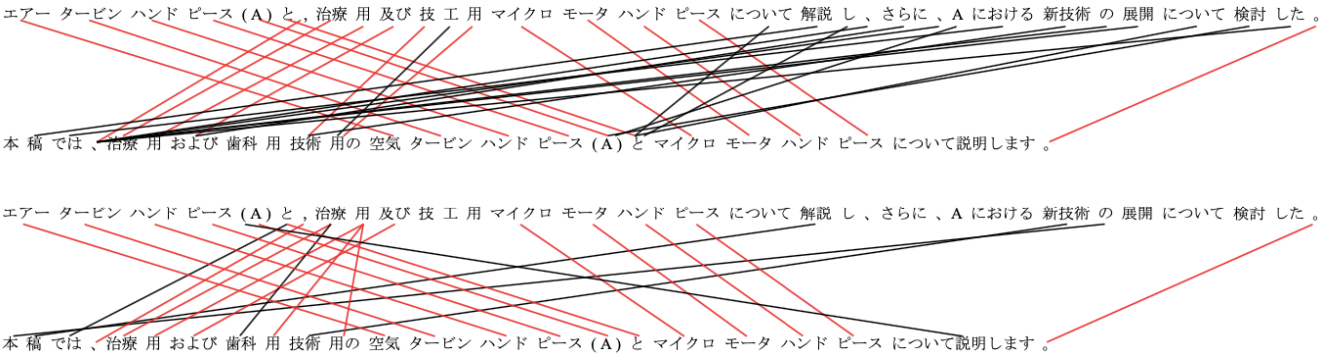

図 A.1　BERTscore の Recall(上)と Precision(下)によるアラインメント(赤が正解、黒が不正解を示す)

ここでは表 4 の 3 行目の BLEU_Sent + RIBES_MeCab + BERTscore + MER_Sent の組み合わせに着目する。表 A.1 に BLEU_Sent のみで見分けることができず、他の評価指標を追加して見分けることができるようになった文の例を示す。

表 A.1 のリファレンスは大まかなパーツに分けると、「解説し」「検討した」となる。DeepL では、少々他の語彙の言い換えがあるものの、色をつけた順番通りに翻訳されており、意味の通じるものになっている。一方、Google 翻訳は「解説し」の部分は訳されているものの、「説明します」と完結しており、「検討した」の部分を翻訳できていない。よって、今回の文は Google 翻訳が一部無視しているので、評価には少し差が出るはずである。RIBES_MeCab、BERTscore、MER_Sent を見てみると、どれも DeepL の方に高評価をしているが、BLEU_Sent はほとんど同じ評価になっており、区別をすることができていない。

それぞれの評価指標の値について表を見ながら考察する。RIBES_MeCab、BERTscore、MER_Sent について、これらが Google 翻訳と DeepL に差をつけることができているのは上記の欠損部分に起因しており、RIBES_MeCab では前半部分の評価値に欠損部分の評価値を加えるか否かで、MER_Sent では欠損部分の挿入操作の増加により、差をつけて評価することができていると考えられる。BERTscore は両者の区別ができているものの、その差が 5%と若干小さい。これは最終的な評価に F 値を用いているためであると考えられる。図 A.1 に BERTscore の Precision と Recall による単語アラインメントを示す。Precision はある程度適切なアラインメントができているが、Recall は欠損している後半部分が無理やり前半部分へとアラインメントされているため、最終的に Recall 値が上がってしまい、Google 翻訳と DeepL での差があまり大きくならなかったと考えられる。それでも、両者をある程度区別することはできている。