システム訳文のみを用いた自動評価との比較による機械翻訳自動評価の分析

高橋 洸丞 [1]　須藤 克仁 [1,2]　中村 哲 [1]
[1] 奈良先端科学技術大学院大学　[2] 科学技術振興機構さきがけ
{takahashi.kosuke.th0, sudoh, s-nakamura}@is.naist.jp

1 はじめに

機械翻訳システムの翻訳能力は年々向上しており, それに伴い自動評価への需要も高まっている. そして, 近年の自動評価の研究も単語やサブワードの意味の近さを評価基準とすることで, WMT metrics task [1, 2, 3] にて評価性能の向上が報告されている. 現在主流となっているのは大規模な事前学習モデルである BERT [4] により参照訳文とシステム訳文を符号化したベクトルを用いて評価値を算出するものである [5, 6]. また我々は入力文を追加で利用し人手評価との相関をさらに改善できることを示した [7].

我々の研究 [7] を通じて, これらの手法は品質の低いシステム訳文では品質の高いものよりも人手評価との相関が低くなることが明らかになっている. ここで主に用いられる自動評価実験用のコーパスに着目してみると, WMT metrics task [1, 2, 3] で提供される人手評価は, DA (Direct Assessment) [8] と呼ばれる 0〜100 の評価値を標準化した実数値を取る. DA スコアは対応する参照訳文に対してシステム訳文の意味的な正確性, また流暢性を共に考慮して一文ずつスコアが与えられている. ここで翻訳誤りには否定関係が逆になっていたり, 副詞や形容詞の係り受けによる意味の違いなど多種の誤りがあり, 実用的な運用の際には重大な誤解を生むリスクのある誤訳に低い評価値を付けるべきである. しかしながら, DA スコアは 1 次元の実数値のみでスコアリングされているため, 文中にどのような翻訳誤りが含まれるのかを直接判別することはできない.

このような背景を踏まえて, 須藤 [9] は, システム訳文の解釈性と正確性をそれぞれレベル分けするように人手評価データを新たに構築した. しかし, BERT regressor [5] のモデルを利用した分類評価実験では, 実務翻訳の立場上許されてはいけないクラスとその他のクラスでの誤分類が見られ, 依然として低品質なシステム訳文の評価が難しいことが示唆される.

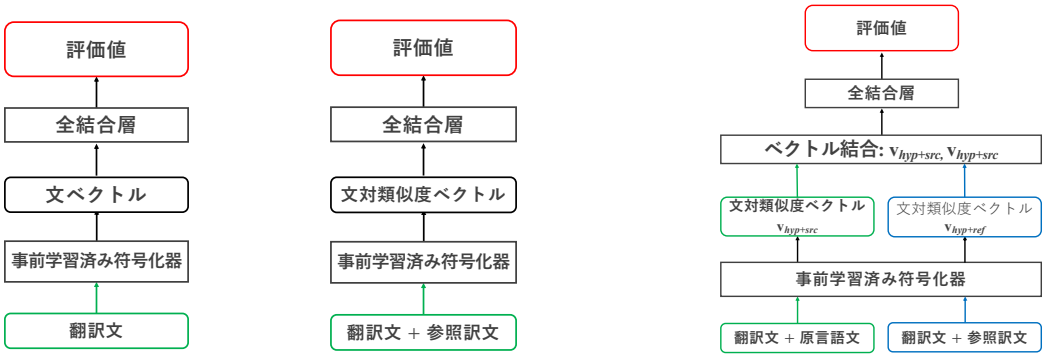
須藤 [9] はコーパス面から, 意味の違いに敏感な評価手法へのアプローチをしたが, そもそも現在主流の評価モデルはどれほど参照訳文とシステム訳文の意味の違いを識別できるのだろうか. 翻訳の評価は大きく分けると, 意味の正確性と流暢性という二つの基準で行われるはずである. 自動評価ではシステム訳文と参照訳文を利用して両者を評価している. 本研究では, 参照訳文を用いずシステム訳文のみを入力とする自動評価, すなわち流暢性は評価できても意味の正確性は評価し得ない方法との比較を通じて, どれだけ意味の正確性を評価できているのかを検証する.

2 評価モデル

評価モデルは, 全て事前学習済みの大規模言語モデルと一層の全結合層から構成される (図 1). システム訳文のみを入力とするモデルは図 1(a) に示す通りで, システム訳文 (翻訳文:hyp) を “[CLS] hyp [SEP]” という形式で入力し, 最終層の [CLS] トークンに相当するベクトルを文ベクトルとみなして, 最終的に全結合層で回帰問題として評価値を計算する. BERT regressor や BLEURT は, 図 1(b) の構造を取り, 翻訳文と参照訳文を [SEP] トークンで区切り, “[CLS] hyp [SEP] ref [SEP]” という形で二文をまとめて入力する.

また我々が提案したモデル [7] は図 1(c) のように, 翻訳文と参照訳文の入力だけでなく, 翻訳文と原言語文をまとめた入力を追加している. これら三つのモデルを以降それぞれ “hyp only”, “hyp+ref”, “hyp+src/hyp+ref” と記述する.

どのモデルも平均二乗誤差 (Mean Squared Error: MSE) を損失関数とし, 誤差逆伝播法により最後の全結合層と符号器の両方のパラメータを更新する.



(a) システム訳文のみ (hyp only)

(b) システム訳文+参照訳文 (hyp+ref)

(c) システム訳文+原言語文/システム訳文+参照訳文 (hyp+src/hyp+ref)

図 1　入力別のモデル図

3　実験

本研究では, システム訳文のみを入力とした意味的な正しさを考慮しない評価モデルにより, どの程度 DA スコアを予測し, 人手評価との相関が得られるのかを, 参照訳文や原言語文を用いた評価手法と比較して検証した.

3.1　実験設定

評価実験は WMT17 metrics task コーパスの all-en で行った. 含まれる言語対は, チェコ語 (cs)-英語 (en), ドイツ語 (de)-英語 (en), フィンランド語 (fi)-英語 (en), ラトビア語 (lv)-英語 (en), ルーマニア語 (ro)-英語 (en), ロシア語 (ru)-英語 (En), トルコ語 (tr)-英語 (en), 中国語 (zh)-英語 (en) の 8 言語対である. その中でも WMT15, 16 に相当し重複を除いた 5344 文の内, 534 文を開発用, 残りの 4810 文を訓練用とした. 評価テストは WMT17 に相当する 3920 文に対して行った.

使用した事前学習済み符号化モデルは RoBERTa [10], 合成データで事前学習した BLEURT [6], XLM-RoBERTa [11] である. 実装は全て HuggingFace Transformers[1)] で行い, RoBERTa は roberta-large, XLM-RoBERTa は xlm-roberta-large を使用した. BLEURT のモデルは BLEURT-large-warmedup を pytorch 用に変換した後に HuggingFace Transformers の BERT 用のクラスで読み込み使用した. RoBERTa と BLEURT モデルは英語のコーパスのみで事前学習されたモノリンガルモデルなので, hyp+src/hyp+ref にはマルチリンガルモデルである XLM-RoBERTa で実験を行った. また参考実験として, システム訳文と原言語文 (hyp+src) を入力としたモデルでの評価実験も XLM-RoBERTa で行った. そして評価性能の比較対象として, 文レベルでスコアを計算した BLEU (sacre BLEU [12]) と BERTscore [13] でも評価実験を行った.

学習時の最適化は Adam [14] で, ミニバッチ学習を最大 10 エポック行った. 各エポック毎に訓練データと開発データはシャッフルし, 開発データのロス値が下がらない場合に学習率を $\frac{1}{\sqrt{2}}$ 倍した. その他のハイパーパラメータは以下の組み合わせの中で, 開発データにおいて人手評価とのピアソンの相関係数値が最も高くなるモデルを評価テストに使用した. またそれぞれの試行毎の初期化によって結果が変動するので, 試行回数もハイパーパラメータとみなしている.

- 初期学習率 : {9e-6, 6e-6, 3e-6}
- ミニバッチサイズ : {4, 8}
- 学習エポック数 : {1, ...,10}
- 最終結合層のドロップアウト率: {0.1, 0.2, 0.3}
- 試行回数 : {1, ..., 10}

本研究の評価実験では, モデルの評価値と人手評価とのピアソンの積率相関係数とケンドールの順位相関係数で評価性能を測る. どちらの相関係数値も −1.0〜1.0 の値を取り, 1.0 に近い程モデルの評価性能が高いことを示す.

3.2　実験結果

表 1, 2 に各モデルの評価値と人手評価とのピアソンの相関係数, ケンドールの相関係数を示す. どの hyp only のモデルも hyp+ref には及ばないが, ピアソンの相関係数が 0.6 を上回り, ケンドールの相関係数が 0.45 前後であった. また全ての hyp only が

1)　https://github.com/huggingface/transformers

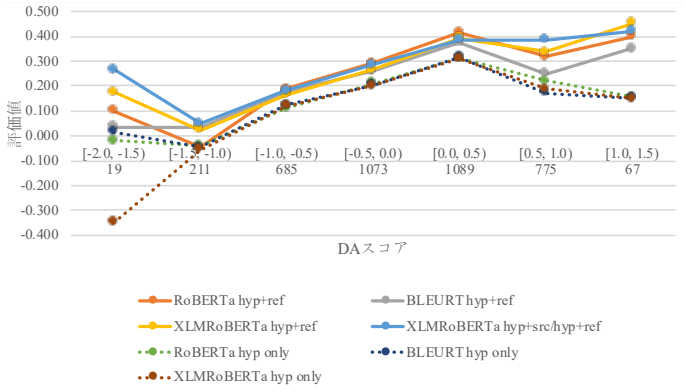


図 2　DA スコアの区間ごとのピアソンの相関係数の関係
（各区間表記の下の数値は区間内のサンプル数を示す）

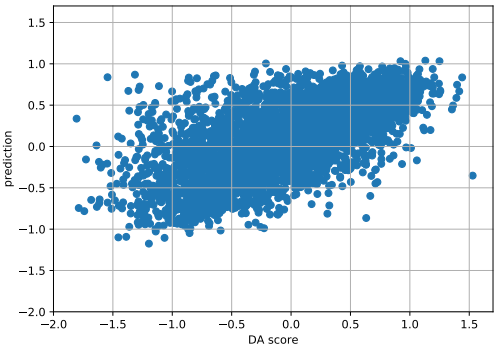


図 3　RoBERTa hyp only による評価予測値と DA スコアの散布図

BLEU を超えており, このコーパスにおいては BLEU などの字句ベースの評価指標より, システム訳文の流暢性にのみ焦点を当てた評価の方が高い相関を示した.

また, 表 3 に hyp+ref の hyp only と比較した際の相関係数の上昇率を示す. hyp+ref は hyp only と比較すると, どの事前学習モデルにおいても, ピアソンの相関係数で約 1.25 倍, ケンドールの相関係数が約 1.3 倍となり, 現在主流な評価モデルは流暢性だけでなく意味の評価がある程度可能だとわかる.

3.3　分析および考察

我々の先行研究 [7] で, hyp+ref の評価モデルは品質の低い翻訳文に対して評価性能が下がることがわかっているが, hyp only ではどうなるのだろうか. DA スコアを 0.5 ずつの一定区間に分けてピアソンの相関係数値を見ることで分析を行った (図 2). その結果, hyp only は $[-2.0, -1.5)$, $[1.0, 1.5)$ の二区間で hyp+ref より評価性能が下がっていることが分かる. それぞれの区域での RoBERTa hyp only の評価予測値と DA スコアの関係を散布図として, 図 3 に示す. この図から, hyp only では低品質な翻訳文で評価値を DA スコアよりも高く予測しているものが多く見受けられ, 高品質なものでは評価予測値が低く出てしまっていることがわかる. 流暢性のみに着目すると, 流暢だが意味的な誤りを含む文や, 流暢かつ意味も正しい文を上手く評価できないことが示唆される. したがって, システム訳文が参照訳文にほぼ等しい時や重大な誤訳を含むとき, 流暢性ではなく意味の正確性を考慮した評価が必要であると考えられる.

現状の hyp+ref 評価モデルでは, 高品質なシステム訳に対して意味の等価性を識別できているが, 中品質, とりわけ低品質な翻訳文において参照訳文との意味の違いを正しく識別できていないと考察できる. 一方で hyp+src/hyp+ref 評価モデルは, 低品質なシステム訳文に対して hyp+ref や hyp only より良い評価性能を記録した.

ケンドールの相関係数値についても同様の結果であったため付録への記載に留める (図 4).

4　おわりに

本研究では, システム訳文のみを入力とした流暢性だけを評価するモデルと, 現在主流の正解文を用いる評価モデルを比較することで, 現在主流のモデルの評価能力を分析した. 分析結果から, 現在主流の hyp+ref モデルは意味の評価がある程度可能ではあるが, 低品質な翻訳文の評価に適さないことがわかった. hyp+src/hyp+ref についても, hyp+ref よりは低品質な翻訳文に対しても評価性能が高いが, 決して十分な評価能力ではないため, より意味の違いに敏感な手法が求められる. また, BLEU を自動評価指標として使用するよりも, システム訳文のみを利用した手法が人手評価と高い相関を示したことから, 特定のデータセットにおける人手評価との相関の高さが意味の正しさの評価における優位性を必ずしも示さないことが明らかになった.

5　謝辞

本研究は JST さきがけ (JPMJPR1856) の支援を受けたものである.

表 1　各モデルの WMT17 metrics task コーパス (all-en) におけるピアソンの相関係数値

model	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	avg	all
sacre BLEU	0.362	0.372	0.505	0.334	0.438	0.479	0.487	0.425	0.418
BERTscore	0.710	0.745	0.833	0.756	0.746	0.751	0.775	0.759	0.718
RoBERTa hyp+ref	0.807	0.808	0.875	0.847	0.846	0.852	0.823	0.837	0.812
BLEURT hyp+ref	0.718	0.724	0.849	0.812	0.785	0.787	0.748	0.775	0.756
XLMRoBERTa hyp+ref	0.751	0.795	0.837	0.842	0.834	0.827	0.728	0.781	0.802
XLMRoBERTa hyp+src	0.672	0.712	0.774	0.793	0.725	0.755	0.658	0.721	0.727
XLMRoBERTa hyp+src/hyp+ref	0.786	0.821	0.847	0.855	0.840	0.836	0.781	0.824	0.812
RoBERTa hyp only	0.623	0.662	0.743	0.679	0.644	0.670	0.604	0.661	0.647
BLEURT hyp only	0.548	0.563	0.752	0.679	0.619	0.654	0.576	0.627	0.615
XLMRoBERTa hyp only	0.574	0.652	0.702	0.703	0.681	0.674	0.588	0.653	0.640

表 2　各モデルの WMT17 metrics task コーパス (all-en) におけるケンドールの相関係数値

model	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	avg	all
sacre BLEU	0.250	0.259	0.344	0.215	0.318	0.319	0.343	0.292	0.293
BERTscore	0.522	0.554	0.646	0.555	0.555	0.569	0.568	0.567	0.534
RoBERTa hyp+ref	0.605	0.627	0.677	0.669	0.638	0.652	0.612	0.640	0.615
BLEURT hyp+ref	0.554	0.550	0.673	0.631	0.601	0.594	0.581	0.598	0.578
XLMRoBERTa hyp+ref	0.588	0.611	0.648	0.637	0.640	0.619	0.582	0.618	0.597
XLMRoBERTa hyp+src	0.498	0.550	0.594	0.585	0.535	0.583	0.481	0.547	0.540
XLMRoBERTa hyp+src/hyp+ref	0.597	0.628	0.657	0.665	0.623	0.634	0.585	0.627	0.616
RoBERTa hyp only	0.457	0.477	0.559	0.488	0.464	0.492	0.436	0.482	0.473
BLEURT hyp only	0.410	0.398	0.559	0.486	0.442	0.480	0.399	0.453	0.447
XLMRoBERTa hyp only	0.422	0.475	0.518	0.504	0.483	0.490	0.414	0.472	0.464

表 3　hyp only と比較したときの hyp+ref の相関係数値の上昇率 (%)
ピアソンの相関係数の上昇率 / ケンドールの相関係数の上昇率

model	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	avg	all
RoBERTa hyp only	128 / 132	123 / 109	117 / 121	125 / 137	130 / 138	127 / 133	134 / 140	126 / 133	125 / 130
BLEURT hyp only	133 / 135	133 / 138	115 / 120	121 / 130	130 / 136	123 / 124	137 / 146	127 / 132	126 / 129
XLMRoBERTa hyp only	131 / 139	122 / 129	119 / 125	120 / 126	122 / 133	123 / 126	124 / 141	120 / 131	125 / 129

参考文献

[1]Ondřej Bojar, Yvette Graham, and Amir Kamran. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pp. 489–513, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[2]Qingsong Ma, Ondřej Bojar, and Yvette Graham. Results of the WMT18 Metrics Shared Task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 671–688, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[3]Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 62–90, Florence, Italy, August 2019. Association for Computational Linguistics.

[4]Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[5]嶋中宏希, 梶原智之, 小町守. 事前学習された文の分散表現を用いた機械翻訳の自動評価. 第 26 巻, pp. 613–634, 9 2019.

[6]Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics.

[7]Kosuke Takahashi, Katsuhito Sudoh, and Satoshi Nakamura. Automatic machine translation evaluation using source language inputs and cross-lingual language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3553–3558, Online, July 2020. Association for Computational Linguistics.

[8]Yvette Graham, Timothy Baldwin, and Nitika Mathur. Accurate Evaluation of Segment-level Machine Translation Metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1183–1191, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[9]須藤克仁, 高橋洸丞, 中村哲. 深刻な誤訳の識別に向けた分類型翻訳評価データセットの構築. 言語処理学会第 27 回年次大会, 2021.

[10]Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[11]Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics.

[12]Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[13]Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

[14]Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *the Third International Conference on Learning Representations*, 2015.

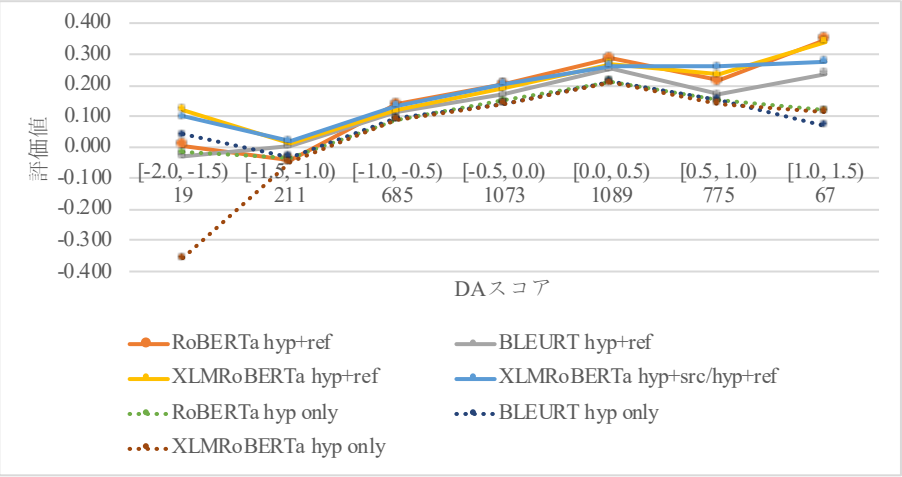


図 4　DA スコアの区間ごとのケンドールの相関係数の関係　（各区間表記の下の数値は区間内のサンプル数を示す）