# ニューラル分類器の予測の解釈に基づく翻訳が難しい表現の検出

坂口 典三　村脇 有吾　Chenhui Chu　黒橋 禎夫
京都大学大学院情報学研究科
{n-sakaguchi, murawaki, chu, kuro}@nlp.ist.i.kyoto-u.ac.jp

## 概要

修辞技法に代表されるような文化圏に特有な表現は近年のニューラル翻訳モデルでも正しく翻訳するのが難しく，原言語側で検出し，前編集することが翻訳精度向上のための有望な方法であると考えられる．そこで本研究では，日本語において翻訳が難しい表現を翻訳前に検出することを試みる．具体的には，翻訳が難しい表現は母語話者が書いたテキストに特徴的であるという仮説を立て，日本語テキストが機械翻訳されたものかそうでないかを予測するニューラル分類器を訓練し，分類器の予測に寄与した表現を翻訳が難しい表現として検出する．実験により，仮説の妥当性を裏付けるとともに，提案手法が翻訳が難しい表現を検出できることを確認した．

## 1　はじめに

異なる言語間には語彙や文法だけでなく，その言語が用いられている文化圏に起因する文化差が表出することがある．例えば日本では「このお店は美味しい」という表現は一般的に用いられているが，Honna [1] によると英語圏では一般的に “restaurant” を “delicious” で形容することはない．この差異はメトニミーにおける指示対象の対応関係が日本と英語圏で異なっていることに起因しており，さらにはそれぞれの文化によって生み出された差異だと言える．

近年ニューラル言語モデルの登場により，機械翻訳は急速な進化を遂げ，従来訳すのが難しかった慣用句などは正しく訳される場合が増えている．．しかし，現在最も一般的に用いられている翻訳サービスである DeepL[1)] でも「このお店は美味しい.」を “This restaurant is delicious” と訳してしまう．この例が示すように，依然として機械翻訳では正しく翻訳することが難しい文化差が存在する．コミュニケーションにおける機械翻訳の実利用を考えたとき，このような文化差を乗り越えるためには，原言語側もしくは目的言語側で適応を行うという方向性が有望であると考えられる．例えば，翻訳が難しい原言語側の表現が翻訳前に検出できれば，その部分を原言語側で翻訳しやすい表現に変えることで翻訳精度の向上が期待できる (表 1).

| 日本語文 | DeepL による英訳 |
|---|---|
| このお店は美味しい. | This restaurant is delicious. |
| このお店の料理は美味しい. | The food at this restaurant is delicious. |

**表 1**　前編集による訳文の変化．上段が原文で下段は下線部が前編集されている．

本研究ではニューラル分類器の予測の解釈に基づいて，翻訳が難しい表現を検出する手法を提案する．提案手法の概略を図 1 に示す．日英翻訳を想定したとき，反対方向の英日翻訳器を用いることで，機械翻訳によって生成された日本語テキストを準備する．そして，日本語母語話者が書いた日本語テキスト (以下，**日本語の原文**) をこの訳文 (日本語) テキストと対照する．キーとなる仮説は，翻訳が難しい表現は日本語の原文に特徴的であるというものである．そこで，まずこの 2 種類のテキストを識別するニューラル分類器を訓練し，次に，分類器の予測に対する入力テキスト中の特定の部分の寄与を明らかにする手法（以下，**説明手法**）を用いて日本語の原文という分類器の予測に最も寄与した部分を翻訳が難しい表現として検出する．

上記の仮説を検証するために，分類器のスコアと日英翻訳の精度の関係を調べたところ，分類器がより機械翻訳らしくないと予測した文ほど実際に翻訳精度が低かった．また，提案手法によって検出された表現を人手で分析したところ，翻訳が難しい表現を見つけられていることを確認した．

1)　https://www.deepl.com/translator

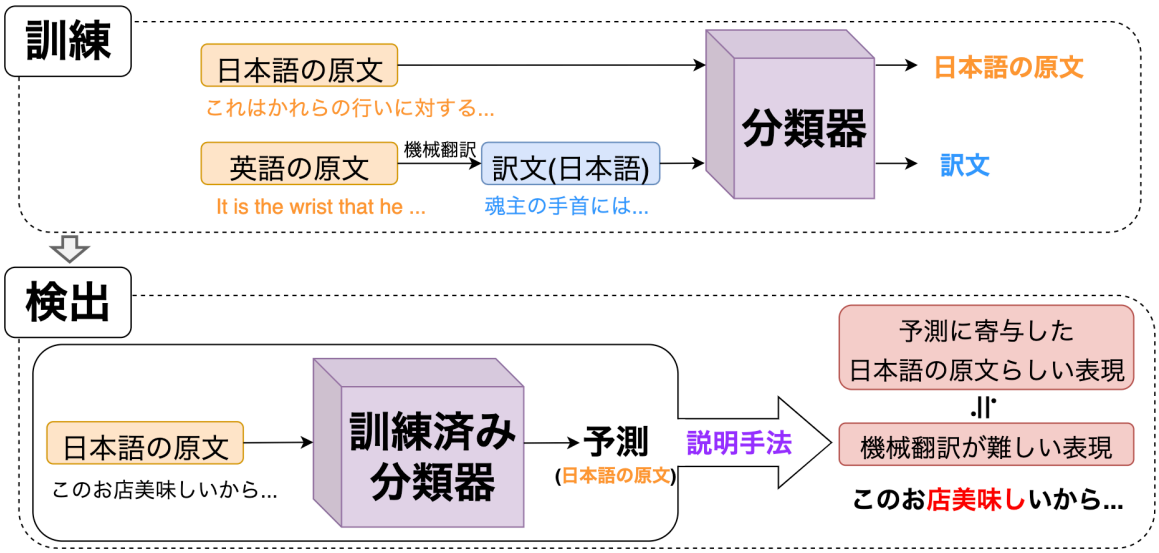


**図 1**　提案手法の概略

## 2　提案手法

本節では図 1 の下段において説明手法を用いる際に行う工夫を本研究の提案手法として紹介する．

### 2.1　Continuously Relaxed Contextual Decomposition

提案手法は，翻訳が難しい表現の検出のためにニューラル分類器の説明手法を用いる．分類器の説明手法は，分類器の予測に対して入力データ中のどの部分が強く貢献したかを明らかにする．提案手法のベースとなる contextual decomposition (**CD**) [2] は，分類器のスコアを入力テキスト中の任意のフレーズによる寄与とその他の部分による寄与に分解する．Murdoch ら [2] は CD を LSTM [3] 分類器に適用したが，その後 Jin ら [4] がこの手法を BERT [5] に拡張している．本研究では BERT を用いる．

CD は，分類器 $y = f(x)$ が $y = g_L(g_{L-1}(...g_1(x)))$ のように操作の再帰的適用で表せることに着目し，各操作 $g(x)$ に対して，$g^{\mathrm{CD}}(x) = (\beta(x), \gamma(x))$ (ただし $\beta(x) + \gamma(x) = g(x)$) となるような近似的な分解を定義する．ここで，$\beta$ は注目する入力内のフレーズの寄与分，$\gamma$ を残りの寄与分である．入力埋め込み層に対して，トークン $t_i$ が注目するフレーズに含まれるなら $(\beta(e_i) = e_i, \gamma(e_i) = 0)$，含まれないなら $(\beta(e_i) = 0, \gamma(e_i) = e_i)$ とし，分解操作を再帰的に適用すると，出力の予測スコア $y$ の分解が得られる．

CD を利用するには入力テキスト中で注目するフレーズを決める必要があるが，トークン列長が $n$ の入力テキスト $T = t_1, \cdots, t_n$ の全てのトークンの組合せ (ただし不連続性を許容) を考慮するには $2^n$ 通りの CD の計算が必要となる．これは一定以上の長さの文では計算量の観点から現実的ではない．

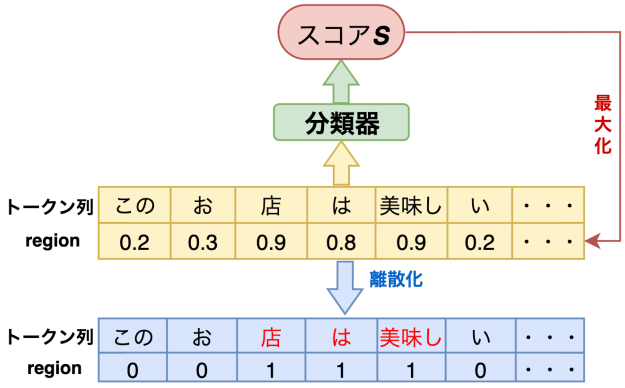


**図 2**　CRCD の概略

そこで $T$ の全てのトークンについて，注目するか否かの離散値ではなく，0 から 1 の連続値 (**region**) に連続緩和する手法 **Continuously Relaxed Contextual Decomposition (CRCD)** を提案する．CRCD の概略を図 2 に示す．CRCD は，分解操作が適用されたネットワーク $g_L^{\mathrm{CD}}(g_{L-1}^{\mathrm{CD}}(...g_1^{\mathrm{CD}}(x)))$ も微分可能であることに着目し，予測スコアの $\beta$ を最大化するような region を逆伝播を使った繰り返し計算により探索する．最後に 0 と 1 に離散化することで，region が 1 となっているトークン列を入力テキスト中で最もその分類らしい表現として検出する．

### 2.2　ニューラルモデル+BoW

分類器は単純に分類損失最小化を目指すだけであり，日本語の原文と訳文 (日本語) のドメインの違い

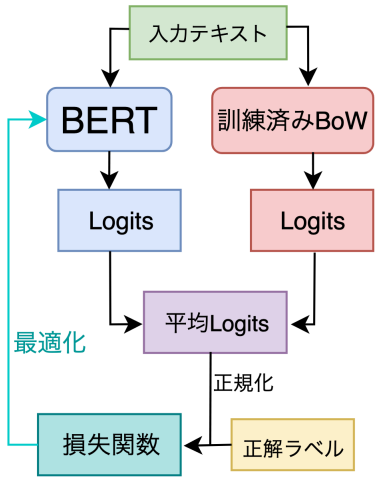


**図 3**　+BoW の概略

といった本質的ではない手がかりを使ってしまう危険性がある．この問題を緩和するために，単語レベルの手がかりに頼る非力な bag-of-words (**BoW**) 分類器を補助的に用いる手法 (+BoW) を提案する．

図 3 に手法の概略を示す．まず bag-of-words を用いた分類器を単体で訓練する．次に BERT を訓練するが，ここで先に訓練していた BoW 分類器が出力するロジットを利用する．訓練用のテキストを入力した際の BERT が出力するロジットを $l_{\mathrm{BERT}}$，BoW 分類器が出力するロジット $l_{\mathrm{BoW}}$ とする．この時 BERT の学習に用いる損失関数 loss は以下のように計算する．

$$\mathrm{loss} = \mathrm{H}(\mathrm{expit}((l_{\mathrm{BERT}} + l_{\mathrm{BoW}})/2),\ R) \qquad (1)$$

ここで H は交差エントロピー誤差，expit はロジスティック関数，$R$ は正解データを表す．上式で計算された損失関数を用いて BERT のパラメータのみを更新する．BoW 分類器が単語レベルの手がかりに基づくロジット $l_{\mathrm{BoW}}$ を提供するため，単語レベルの手がかりだけで分類できるテキストに対しては，BERT のパラメータはあまり更新されない．単語レベルの手がかりだけでは分類できなようなテキストが入力された時，BERT のパラメータは大きく更新される．こうして，BERT が単語レベルよりも深い手がかりに注力するように誘導する．

# 3　実験

本節ではデータセットを構築し，そのデータセットを用いて分類器を訓練する．さらに検出方法の前提となる仮説を検証した上で，実際に翻訳が難しい表現が検出できるか確認する．

## 3.1　データセット構築

データとして，日本語の原文，英語の原文，英語から機械翻訳された訳文 (日本語) が必要となる．この訳文生成のためには英日翻訳モデルも訓練しなければならない．

まず Wikipedia の本文から日英テキストを抽出した．ただし，複数の言語版が存在する一般名詞を扱うページのみを使用した．これはテキスト中に登場する単語に日本語版のページにしか登場しないような固有名詞が存在すると，分類器がそれに強く影響される恐れがあるからである．一般名詞の抽出には森羅プロジェクト[2)]の成果を利用した．森羅プロジェクトは Wikipedia 日本語版のタイトルに拡張固有表現を付与しており，階層ラベル 0 が一般名詞を表している．さらに各ページに記載されている他言語版へのリンク数が 35 以上となっているページを抽出した．こうして抽出した英語版ページと日本語版ページの本文を文単位に分割し，1,073,431 万文の英文と 648,507 万文の日本語文が得られた．

次に英日翻訳モデルを訓練した．Morishita ら [6] が公開している事前学習モデルを WikiMatrix [7] の日英対訳データを用いて追加学習した．学習用データとして 479,315 文，テスト用データとして 1,000 文の対訳データを用いた．訓練時のハイパーパラメータは Morishita らの論文に基づく．

訓練したモデルを SacreBLEU [8] で精度を評価すると 21.82 となった．同じテストデータで DeepL の英日翻訳精度を評価すると SacreBLEU は 16.75 であり，訓練したモデルの精度が上回った．追加学習により Wikipedia の文体にモデルが特化したためであると考えられる．このモデルを用いて英文を翻訳することで，原文とあわせて 1,721,938 文の日本語データが得られた．このデータ内の各文に対しては，機械翻訳で生成された文か日本語の原文かを示すラベルが付されている．

また，後述の分析のために日英翻訳モデルを英日翻訳モデルと同じ設定で訓練した．上述のテストデータで評価したところ，SacreBLEU は 27.76 となった．

## 3.2　分類器の訓練

3.1 節で構築した日本語データを用いて日本語の原文か訳文か予測する分類器を訓練した．1,000 文

2)　http://shinra-project.info/

をテストデータとして抽出し，残りのデータを訓練用データとして用いた．分類器のモデルとして BERT と BoW を用いた．

BERT としては NICT が公開している事前学習モデル[3)]を使用した．また，BoW モデルは線形層 2 層から構成されたニューラルネットを使用し，入力文中の単語は BERT の埋め込み層を複製してベクトル化した．単体での BERT とは別に，BERT を+BoW を用いて訓練した．

各モデルのテストデータにおける分類精度を表 2 に示す．BoW 以外のモデルはほとんど同じ精度となった．BERT+BoW 設定で訓練した BERT を単体で用いても精度が下がらなかったことから，データセット内の文には単語レベルを超える手がかりが十分に存在していると推測できる．

### 3.3 分類結果と翻訳精度の相関

翻訳が難しい表現は日本語の原文に特異的に出現し，英日翻訳によって生成された訳文には含まれていないという仮説を検証した．もし仮説が妥当であれば，分類器が判断した日本語の原文らしさと日英翻訳精度には負の相関があることが期待される．

3.1 節で機械翻訳モデルのテストに用いた WikiMatrix の対訳データから日本語文 1,000 文を選んで単体で訓練した BERT に入力し，分類スコアを計算した．ここでの分類スコアとは，分類器が出力する 2 つのロジットの差をとったものであり，分類スコアが大きいほど日本語の原文である可能性が高いことを示す．さらに 3.1 節で訓練した機械翻訳モデルで日英翻訳を行い，文単位で翻訳精度を計算した．文単位の翻訳精度の評価指標としては SentBLEU [9] と BLEURT [10] を用いた．

翻訳精度と分類スコアの散布図を表 A.1 に示す．翻訳精度と分類スコアのピアソンの積率相関係数を計算すると，SentBLEU で −0.24，BLEURT で −0.45 となった．SentBLEU は文ごとの n-gram を計算するため，意味的に正しい翻訳文でも不当に低い値を計算することがあり，散布図に示すように多くの文が 0 に近い値をとった．BLEURT は従来の評価指標よりも人間の評価に近いとされており，こちらではより強い負の相関が見られた．この結果は仮説を裏付け，分類器にとっての日本語の原文らしさが日英翻訳を困難にすることを示唆する．

| | BERT | BoW | BERT(+BoW) | BERT+BoW |
|---|---|---|---|---|
| 精度 | 0.94 | 0.88 | 0.93 | 0.93 |

**表 2** 分類器のモデルの精度．BERT は BERT 単体で訓練した精度，BERT(+BoW) は+BoW を用いて訓練した BERT モデルを単体で評価した精度，BERT+BoW は+BoW を用いて訓練した BERT のロジットとと BoW のロジットを+BoW と同じ手法で組み合わせて計算した精度を示す．

### 3.4 表現の検出結果の分析

3.3 節の結果から，学習済みの分類器には翻訳が難しい表現に関する情報が含まれていることがわかった．そこで説明手法を用いて文中のどの部分が翻訳が難しい表現なのか検出することを試みた．

まず+BoW の有効性が示された例を表 A.1 に示す．BERT のみの結果では「成敗」や「禁令」のような日本語の原文というドメインに特有の語彙が検出されているが，このような表現は機械翻訳で生成されることが稀な単語であり，翻訳が難しい表現ではない．それに対して BERT+BoW ではある程度それらの表現に注目しないようになっていることがわかる．

次に，検出された表現で実際に機械翻訳が適切でない翻訳を生成した例を表 A.2 に示す．「いい年になった」や「端から見ると」などの表現が検出された．これらの表現の DeepL による訳文を見ると，“reached a good age” や “from the out side” という不自然な訳になっており，実際に翻訳が難しい表現であることが確認できる．ただし，検出されたフレーズが実際に不自然な翻訳をされているかを自動的に評価する方法が確立されておらず，検出結果の定量的評価が課題として残されている．

## 4 おわりに

本研究では機械では正しく翻訳することが難しい表現は日本語母語話者が書いた日本語テキストに特異的に出現し，機械翻訳によって生成されたテキストには含まれていないという仮説を検証した．さらに，分類器の説明手法によっていくつかの翻訳が難しい表現を検出した．また，bag-of-words 分類器を補助的に用いることで，分類器に単語レベル以上の手がかりに着目させる手法を提案し，その有効性を確認した．今後の課題としては，提案手法の定量的評価方法と前編集の方法の確立が挙げられる．

3) https://alaginrc.nict.go.jp/nict-bert/index.html

## 謝辞

本研究は一部 JSPS 科研費 21K12029 の助成を受けた.

## 参考文献

[1] Nobuyuki Honna. That restaurant is delicious. [japan]. **Asian Englishes**, Vol. 13, No. 2, pp. 64–65, 2010.

[2] W. James Murdoch, Peter J. Liu, and Bin Yu. Beyond word importance: Contextual decomposition to extract interactions from LSTMs. In **Proceedings of 6th International Conference on Learning Representations, ICLR 2018**, 2018.

[3] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. **Neural Computation**, Vol. 9, No. 8, pp. 1735–1780, 1997.

[4] Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. In **International Conference on Learning Representations**, 2020.

[5] 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.

[6] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In **Proceedings of The 12th Language Resources and Evaluation Conference**, pp. 3603–3609, Marseille, France, May 2020. European Language Resources Association.

[7] Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 1351–1361, Online, April 2021. Association for Computational Linguistics.

[8] 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.

[9] 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.

[10] 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.

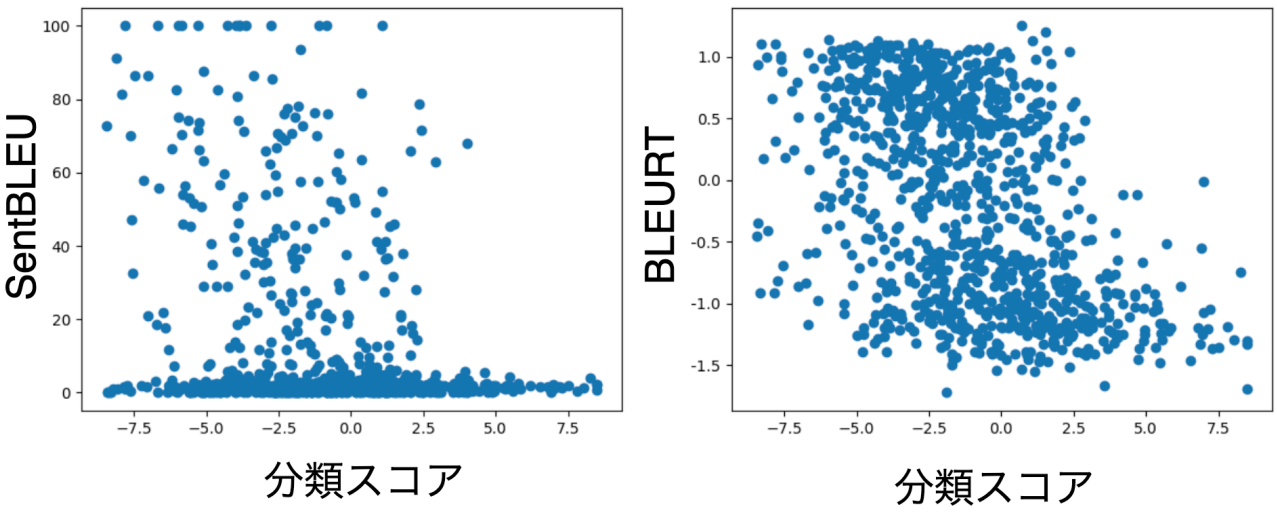


**図 A.1**　翻訳精度と分類スコアの散布図

| モデル | 検出結果 |
|---|---|
| BERT のみ | これ ら の 奴隷 商人 は 下級 の 兵士 と 通じて 朝鮮 人 を 調達 して いた ため 、<br>加藤 清 正 など は 「 乱 暴 [ U N K ] に 身分 の 低い 者 を こき 使う 者 が あった ならば 、<br>その 主人 の 責任 と して 成 敗 を 加える 。 」 と 禁 令 を 発して いる 。 |
| BERT+BoW | これ ら の 奴隷 商人 は 下級 の 兵士 と 通じて 朝鮮 人 を 調達 して いた ため 、<br>加藤 清 正 など は 「 乱 暴 [ U N K ] に 身分 の 低い 者 を こき 使う 者 が あった ならば 、<br>その 主人 の 責任 と して 成 敗 を 加える 。 」 と 禁 令 を 発して いる 。 |

**表 A.1**　+BoW の有効性を示す例．下線部は検出された箇所を示している．

| | |
|---|---|
| 検出結果 | 特に 日本 で は 、 2000 年 代 以降 、 いい 年 に なった 人 達 が 、<br>端 から 見る と 幼稚 で 浅 は か に も 思える 事件 を 起こし 、<br>メディア で 報道 さ れる こと も 増えて きて いる 。 |
| DeepL 訳 | Especially in Japan, since the 2000s, there have been an increasing<br>number of incidents reported in the media by people who have<br>reached a good age that seem childish and shallow from the outside. |

**表 A.2**　翻訳が難しい表現が検出できた例．下線部は上段では検出された箇所を示しており下段では対応する訳を示している．