修士学位論文要約（令和３年３月）

機械翻訳モデルの頑健性評価に向けた言語現象毎データセットの構築と分析

藤井　諒
指導教員：乾　健太郎

Phenomenon-wise Evaluation Dataset Towards Analyzing Robustness of Machine Translation Models

Ryo FUJII
Supervisor: Kentaro INUI

Neural Machine Translation (NMT) has shown drastic improvement in its quality when translating clean input such as text from the news domain. However, existing studies suggest that NMT still struggles with certain kinds of input with considerable noise, such as User-Generated Contents (UGC) on the Internet. To answer the question of what creates the large performance gap between the translation of clean input and that of UGC, we present a new dataset, PheMT, for evaluating robustness of MT systems against specific linguistic phenomena in Japanese-English translation. Our experiments with the dataset revealed that not only our in-house models but even widely used off-the-shelf systems are greatly disturbed by the presence of certain phenomena.

1. はじめに

ニューラル機械翻訳（NMT）の発展により，ニュース記事のように文体の整った入力に対する翻訳は既に人間の翻訳に匹敵するレベルにまで到達したとも言われる．しかし，そのめざましい発展をもってしても，ソーシャルメディアなどのユーザ生成コンテンツ（UGC）に対するNMTの適用可能性は低い[1]．この背景には，翻訳品質の低下を招くUGC特有の要因が不明瞭であり，性能向上の端緒となる頑健性の測定基盤が十分に整備されていないことがあげられる．

そこで本研究では，機械翻訳システムの精緻な評価に向けた第一歩として日英機械翻訳に焦点を当てる．具体的には，UGCに頻出の固有名詞，名詞の省略，口語表現，異表記の4つの言語現象に着目した言語現象毎評価データセットPheMTを構築した（図1）．データセットを用いた評価と分析により，UGCという限られたドメインにおいても，広く商用に利用される機械翻訳システムを含む最先端のNMTモデルが，依然として多くの課題を抱えていることを明らかにし，機械翻訳評価における一つの方向性として言語現象に着目することの有用性を示した．

2. 現象毎データセットの構築

本研究では，UGCにおける機械翻訳システム評価のためのベンチマークであるMTNTデータセット[2]に言語現象のアノテーションを行った．個々の現象データセットが評価に十分な文数からなるよう，PheMTの構築には訓練および開発用データを含む全データを用いることとした．元来評価の用に作成されていないデータを用いることによる品質の低下を防ぐため，翻訳品質の基準となる十分性スコアを定義し，クラウドソーシングによるアノテーションを通して両言語における意味の等価性が担保される高品質な文対のみを抽出した．

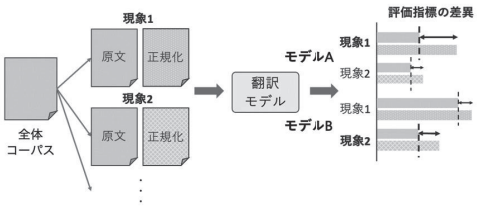

図1 現象毎データセットによる評価．

評価の対象とする言語現象としては，UGCを含む様々なテキストに頻繁に現れ，他言語対においても翻字や転写の文脈で注目を浴びている固有表現に関して，「固有名詞」および「名詞の省略」を定義した．さらに，日本語を対象とする他の言語処理タスクにおいて問題となることが指摘されている崩れ表記に着目し[3]，長音の挿入や発音の崩れなどからなる「口語表現」，代表的な表記と文字種の異なる「異表記」を定義することで計4種類の現象を対象とした．

アノテーションは問題の難易度を考慮し，現象ラベルの付与，該当表現およびアラインメントの抽出，正規化の3つの段階に分けて行った．ここで「正規化」とは，ある現象に分類される表現に対し，その現象に分類される理由を取り除くような変形を適用することを意味する．つまり，名詞の省略に分類される「アプデ」は，これが名詞の省略でなくなるように「アップデート」に正規

化される. なお, 固有名詞に関しては正規化の方法が自明でないため, 表現とアラインメントの抽出のみを行った. 構築したデータセットの一例を表 1 に示す.

3. 翻訳モデル

実験は以下の 3 つの観点から設定した 5 種類のモデルと, Google 翻訳, DeepL 翻訳の 2 つの商用システムに対して行った.

1. 各現象は訓練データの拡充により解決されるか
   - **SMALL**, **LARGE**
2. トークンへの分割手法は頑健性に影響するか
   - **CHAR**
3. 構築したデータセットを用いて, 特定の現象に対する頑健性の向上を検出可能か
   - **PRON**, **CAT**

PRON および CAT の 2 モデルでは, 前処理として形態素解析ツールキット MeCab[4]を用いた発音への転写を行った. この前処理は, 特に異表記に対する頑健性の向上を目的としている. 例えば,「アリガトウ」という表現はカタカナでの出現は一般的でないものの, 発音への変換を経ることで頻出の表現に集約することができる. 推論時に未知の表現が減少し翻訳品質が改善することが期待され, これを構築したデータセットにより定量的に把握可能であるかを検証する.

4. 現象毎評価

評価は MTNT データセット中の原文と正規化後の文をモデルに入力し, 任意の評価指標の差分を用いて行った. また, 同一の差分であればスコアの絶対値が小さい場合により影響が大きいと考えられるため, 差分を正規化後のスコアで除算した頑健性スコア ROBUST を定義した. これは, 正規化後の入力で到達可能な翻訳品質に対する相対的な品質の低下と捉えることが出来る.

$$\mathrm{ROBUST} = \Big(score(x_{orig}, y) - score(x_{norm}, y)\Big) / score(x_{norm}, y)$$

表 2 は現象該当表現の翻訳正解率を評価指標として用いた場合の各モデルの ROBUST スコアである. なお, 正規形を有しない固有名詞に対しては原文入力時の正解率を示す. SMALL と LARGE を比較すると固有名詞では訓練データの拡充により正解率が大幅に向上することが確認できる. また, SMALL の名詞の省略に対する頑健性スコアは他のモデルに比べ著しく低くとどまった. 一方で, 口語表現と異表記の 2 現象においては, 必ずしも訓練データ規模の大きなモデルが優位ではないようである. 特に異表記に対しては, 広く商用に利用される機械翻訳システムをもってしても極めて低いスコアを示した. トークンへの分割手法という観点では LARGE と CHAR の比較により, 文字ベースの分割が口語表現に対する頑健性を向上させることがわかった. これは, 口語表現の正規化前後の編集距離が他の現象に比べ小さく, 周囲の文字が大きな手がかりとなるためであると考えられる. また, 構築したデータセットが, 特定の現象に対する頑健性に感受性を有するかに関して, LARGE と PRON および CAT の 2 モデルを比較すると, 後者で頑健性の向上が期待される異表記に対してスコアが大きく改善することが確認された. この結果から, PheMT が従来の評価には現れない側面に対して再現性の高い評価を提供できること, UGC を取り扱う際には起こりうる言語の変化に対してその特性などから考察を行う事が重要であることを確認した.

表 1 構築したデータセット中の文例.

固有名詞	
Orig. (Ja)	まじで**メルカリ**で野菜売ってるー！
Ref (En)	There really are vegetables for sale at **Mercari**!
異表記	
Orig. (Ja)	他は**しっと**案件
Norm. (Ja)	他は**嫉妬**案件
Ref. (En)	Anything else is just **jealousy**.

表 2 正解率を用いた場合の頑健性スコア.

	SMALL	LARGE	CHAR	PRON	CAT
固有	(34.80)	(48.94)	(47.40)	(43.12)	(47.96)
省略	-15.41	+5.94	+1.23	-0.32	+10.90
口語	-25.85	-33.46	-26.48	-60.76	-52.57
異表記	-63.52	-65.45	-65.19	-26.47	-22.40

5. まとめ

本研究では, ユーザ生成コンテンツに着目した日英機械翻訳の言語現象毎評価データセットを提案した. 構築したデータセットを用いた分析により, UGC という限られたドメインにおいても, 広く商用に利用されるシステムを含む多くのモデルに依然として課題が残されていることを明らかにし, 機械翻訳評価において言語現象に着目することの有用性を示した.

文献

[1] Li et al., Findings of the First Shared Task on Machine Translation Robustness. In WMT, pp. 91-102, 2019.
[2] Michel et al., MTNT: A Testbed for Machine Translation of Noisy Text. In EMNLP, pp. 543-553, 2018.
[3] Sasano et al., A Simple Approach to Unknown Word Processing in Japanese Morphological Analysis. In IJCNLP, pp. 162-170, 2013.
[4] Kudo et al., Applying Conditional Random Fields to Japanese Morphological Analysis. In EMNLP, pp. 230-237, 2004.