ユーザ生成コンテンツの高品質な自動翻訳に向けた言語現象の体系的分析

Systematic Analysis of Linguistic Phenomena for Better Understanding Translation on User-Generated Contents

藤井 諒 [*1] 三田 雅人 [*2*1] 阿部 香央莉 [*1*2] 塙 一晃 [*2*1] 森下 睦 [*3] 鈴木 潤 [*1*2] 乾 健太郎 [*1*2]
Ryo Fujii Masato Mita Kaori Abe Kazuaki Hanawa Makoto Morishita Jun Suzuki Kentaro Inui

[*1]東北大学 [*2]理化学研究所 [*3]NTT コミュニケーション科学基礎研究所
Tohoku University RIKEN NTT Communication Science Laboratories

Neural Machine Translation (NMT) has shown drastic improvement on its quality when translating clean input. However, it still struggles with some kind of input with plentiful of noises, like User-Generated Contents (UGC) on the Internet. In order to make NMT systems indeed useful in promoting cross-cultural communication, one of the most promising direction we have to follow is to correctly handle with these input. Though necessary, it is still an open question that what brings the great gap of performance between translation of clean input and UGC. In this paper, we conducted systematic analysis on current dataset focusing on UGC and made it clear which linguistic phenomena greatly affected the translation performance.

1. はじめに

ニューラル機械翻訳（NMT） [1] の登場により，高品質かつ大規模な学習データを入手可能なニュース記事や，出現語彙が限定された旅行会話などの翻訳品質は著しく向上した．しかし，ソーシャル・ネットワーキング・サービス（SNS）に代表されるユーザ生成コンテンツ（UGC）を対象とした場合，現状の NMT システムにおいて十分な翻訳品質を担保するのは難しい．これは，タイポや誤変換，ユーザによる意図的な表層の改変を含む表記揺れ，日々新たに生成される固有名詞やネットスラングなどといった，従来の翻訳コーパスには稀有な言語現象の存在に起因すると考えられる．このような言語現象を含む UGC を適切に自動翻訳できるようにすることは，機械翻訳システムを異文化・他言語交流といった場で真に使えるコミュニケーションツールとする上で必須の要件であり，次に取り組むべき大きな課題の一つであると言える．一方で，NMT システムがどのような言語現象の存在に大きく影響されるのかは明らかでなく，偏在するユーザコンテンツを適切に翻訳できるシステムの構築に向けての方向性は依然定まっていない．

そこで本研究では，UGC に頻出の言語現象が NMT システムに与える影響について調査した．具体的には，まず言語現象の類型化を行い，各言語現象に対応するアノテーションラベルを設計した．次に，設計したアノテーションラベルに基づき，現象箇所のみを訂正することで，各現象の有無のみが異なる評価用データを作成した．これにより，1) 現象毎の影響をより適切に評価できるようにデータセットを拡充し，2) 意図的な「かな表記」が学習データの増加により解消されない，特別な対処を必要とする現象であることを明らかにした．

2. 関連研究

UGC に対する翻訳精度のベンチマークとして，Michel ら [2] はオンラインディスカッションサイト Reddit[*1] からクロールした一連のコメントに対し，プロの翻訳家による対訳を付与することで，MTNT データセットを作成した．Michel らは，データセットの分析により MTNT が従来のニュース翻訳などのコーパスに比べ，タイポ・文法誤りなどのノイズを多く含むことを明らかにした．一般に，NMT は従来の統計的機械翻訳 （SMT）に比べノイズに対する感受性が高く，たとえノイズが少量でも翻訳の品質を大きく悪化させることが知られている [3]．こうした背景から，近年ノイズに頑健な機械翻訳システムの構築を目的としたタスクも開催されるようになった [*2]．

しかし我々は，現状このタスクにおける「ノイズ」には，部分的な翻訳・ミスアラインメントなどの「文対としてのノイズ」と，文法誤り・スラングを含む「文内のノイズ」の多義性があると考えている．「文対としてのノイズ」に対しては，コーパスフィルタリングの有効性が示されており，上述のタスクにおける精度向上にも大きく寄与した [4, 5]．一方で， フィルタリングに基づくアプローチはドメイン非依存的であり，UGC という特異なドメインにおけるシステムの優位性を帰結する根拠としては不十分である．「文内のノイズ」の影響に注目した研究として，Anastasopoulos ら [6] は，特に文法誤りに注目して NMT システムの頑健性向上を試みた．具体的には，学習者コーパスを用いて求めたノイズの分布に従い，評価データおよび訓練データに人工ノイズを付与した評価セットを作成し，ノイズの有無のみが異なる 2 種類の評価データにおける評価を比較することで，人工ノイズを付与した疑似データによる学習が頑健性の向上に有効であることを示した．一方で，今回我々が翻訳対象として焦点を当てる UGC には，文法誤りのみならず，NMT システムの出力を乱しうる多様なノイズが含まれていると考えられる．しかし，言語現象に着目した NMT システムの分析は未だ十分に行われておらず，どのような現象の存在がニュースなどの翻訳と UGC の翻訳の間に存在する品質の差異を生み出すのかは依然として明らかになっていない．

3. 手法

本稿では，UGC に対する翻訳精度に影響を及ぼすような言語現象を明らかにすることを試みる．具体的な手法としては，言語現象ラベルの定義とアノテーションを行い，ある言語現象を含む文に対し，その該当箇所を打ち消した評価データセットを構築する．Anastasopoulos らの手法に則り，対になる評価データを用いた評価の比較により各現象の影響度を測定する．

連絡先: 藤井 諒 (Ryo Fujii) 東北大学大学院 情報科学研究科 乾・鈴木研究室
〒 980-8579 宮城県仙台市青葉区荒巻字青葉 6-6-05 東北大学工学研究科 電子情報システム・応物系 1 号館 6 階
E-mail: r-fujii(at)ecei.tohoku.ac.jp
TEL: 022-795-7091

*1 https://www.reddit.com

*2 http://www.statmt.org/wmt19/robustness.html

表 1: 言語現象ラベルの定義と分類例

現象ラベル	定義	分類例
固有名詞	人名，作品名，企業名，ブランド名など	安倍首相，アナと雪の女王
名詞の省略	正式名称の短縮，頭文字を用いた表記など	マスコミ，JK
スラング	ネット上・特定のコミュニティ内でのみ用いる言葉	すこ，w
かな表記	一般的には漢字などで表記されるひらがな，カタカナ表現	とうぜん，キモチワルイ
スタイル変化	辞書的な表記から変化したもので「非正規形」に当てはまらないもの	やで，だにゃん
非正規形	長音・小文字化，母音・子音の変化とその複合	かなちい，だろー
誤変換	タイプミス・変換時のミスによる本来の表記と異なる表記	ふいんき，体学歴
漢字化	一般的にひらがな，カタカナで表記される表現の漢字化	只，迄
複合語	複数の語の組みあわせで構成される語，そのようにして作られた造語	ちょいワル，アベノミクス
特殊文字	記号，unicode 絵文字など	ω，🤔
他言語	翻訳対象言語（日本語）以外で記述された表現	What is this subreddit about?
R-18	成人向け，不適切な表現	（省略）
非文または一文でない	記号のみからなるもの，アスキーアートなど	（省略）

言語現象ラベルの定義

MTNT の学習データおよび開発用データを結合し，全体から無作為に取り出した 500 文を対象に，言語現象の列挙を行った．評価データと同じドメインに属するこれらの文の観察により，我々は 13 種類の言語現象ラベルを定義した．ラベルの定義とその分類例を表 1 に示す．

言語現象ラベルの付与方法

表 1 のラベル定義に従い，MTNT の評価データおよびブラインドデータ [*3] の分類を行った．文中に含まれる現象と該当箇所を複数回答可能な形式で抜き出すタスクを設計し，クラウドソーシングを用いてラベルを付与した．該当箇所の抜き出しの際に生じる，現象に該当する最小単位での抜き出しおよび単語や文節単位での抜き出しの恣意性については，“ある回答が他の回答の一部を含む場合に，重複部分を複数回の出現と見なす” ことで対処した．具体的には，固有名詞の該当部分として「安倍首相」および「安倍」が付与された場合，部分的に包含される「安倍」を 2 回の出現と数えた上で，複数人が付与した最長の表現をラベルの該当部分とした．

現象毎データセットの構築（固有名詞）

ある文から固有名詞の存在を打ち消す方法は自明でない．これに対し，我々は固有名詞部分を「某〇〇」という形の上位語表現に置き換えることで対処した．具体的には，「企業」，「ゲーム」など 16 個の語彙からなる閉じた語彙セットを用意し，固有名詞の該当箇所について本稿の著者 3 名でアノテーションを行った．語彙セットには，固有名詞の網羅率とラベルの排他性，対訳の自明性を重視して定義した以下のものを用いた．

> 彼（he, him），彼女（she, her），氏（person），キャラクター（character），グループ（group），製品（product），食品（food），企業（company），組織（organization），ゲーム（game），アニメ（anime），作品（work），サイト（site），アプリ（app），イベント（event），所（place）

言い換えに伴う対訳の編集に際しては，クラウドソーシングを用い，原言語側の固有名詞に対応する対象言語文中の箇所を同定するアラインメントタスクを行った．その後，対象言語側の対応箇所を “a certain” + その訳語に置き換えることで評価データを作成した（例．“Shibuya” → “a certain place”）．また，一文中に同じ表現による言い換えが複数回出現する文については，現象の評価とは無関係な訳抜けなどが生じる影響を考慮し評価データとして用いなかった．

*3 タスクにおいて，開催時には未公開とされる評価データで順位の決定に用いられる．以降，評価データとブラインドデータの結合データを「評価データ」と言及する．

現象毎データセットの構築（固有名詞以外）

固有名詞以外の現象に対しては，原言語（日本語）側の文とラベルに該当する部分を見せて該当表現の正規化を行うタスクを設計し，クラウドソーシングを用いて評価データを構築した．あるラベルに属する表現の訂正（正規化）は，そのラベルの定義に該当する部分のみとし，訂正後に他の現象に帰着する場合にも，多段階の訂正は行わないようにした．例として，非正規形に該当する表現には，かな表記を複合する場合も多く見られたが（例．かなちい），漢字化による一般的表現への書き換えはタスクのスコープ外とした．これらの手順に従い，表 2 のような評価用データセットを作成した．

4. 実験設定

クラウドソーシング

MTNT データセットの評価データ全 2112 文に対し，対話コーパスのフィルタリングのために作成された用語リスト [*4] を用いて不適切な表現を含む文の除去を行った結果，1895 文をアノテーション対象として得た．これらの文に対し，表 1 に列挙した現象のいずれかを含むかどうか，また含まれる場合どこが該当箇所かを複数記述可能な自由記述方式で解答するタスクを設計した．結果を集計した後，複数人のラベルの一致が見られた箇所に対し，3. 節に示す手順に従って現象毎のアノテーションを行った．回答の品質担保のため，すべてのタスクにチェック設問を設け，1 設問に対しチェック設問を通過した 5 人のワーカの回答を集約した．正規化タスクに関しては，過半数の回答が一致したもののみを以降の評価データとして使用した．これらのタスクは Yahoo!クラウドソーシング [*5] を用いて行った．

モデルアーキテクチャ

すべての実験において，`fairseq` ツールキット [7] の Transformer-base アーキテクチャ [8] を用いた．ハイパーパラメータは Murakami ら [9] の設定に準じた．

学習データ

WMT2019 Robustness Task にて提供された TED，KFTT，JESC，MTNT の 4 種類のデータセット（constrained），および JParacrawl v1.0 [10]（unconstrained）を用いた．JESC コーパスは，英語側がすべて小文字で記述されていたため，JESC を除く 3 種類のタスク提供コーパスを用いて `moses` ツールキットの recaser を学習し，該当部分に適用することで文字種情報の復元を行った．また，Murakami ら [9] の先行研究に倣い，MTNT データセット中の絵文字および顔

*4 `https://github.com/1never/open2ch-dialogue-corpus`
*5 `https://crowdsourcing.yahoo.co.jp`

表 2: 作成したデータセットの一例　　注：{ 置換前/**置換後** }

固有名詞データセット	Ja	去年 { 渋谷/**某所** } で記念展示会やってたね
	En	There was a celebration exhibition held in {Shibuya/**a certain place**} last year.
かな表記データセット	Ja	{ きゅうじつしゅっきん/**休日出勤** }(震え声
	En	I went to work on a day off (shivering voice
名詞の省略データセット	Ja	地味な { アプデ/**アップデート** } だが
	En	That's a plain update though

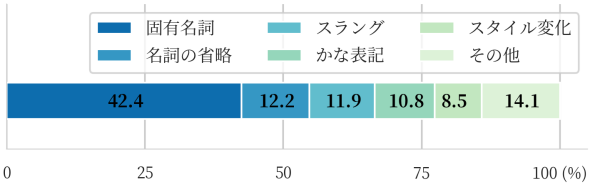

図 1: 各現象の出現頻度

文字にはプレースホルダを適用し，後処理の段階でプレースホルダを元の文字列に書き換える処理を行った．対訳ペアの規模は constrained データが 3.89M，unconstrained データが 12.2M であり，`sentencepiece` を用いて，語彙数が 32,000 となるようにそれぞれのデータで学習した Byte-Pair-Encoding（BPE）[11] を適用した．それぞれのデータで学習した 2 モデル（base）と，それらを MTNT データで finetune した 2 モデル（tuned）の合計 4 モデルを評価対象とした．

評価指標

評価には，自動評価指標 BLEU[12] を用いた．MTNT データセットおよび作成したデータセットに対して翻訳を行い，BLEU の差をもって現象毎の影響度の大きさを測定した．

5. 実験結果

5.1 定量分析

図 1 に，クラウドソーシングにより付与された言語現象ラベルの相対頻度を示す．ここでは各文に最も多く付与された言語現象ラベルをその文の代表ラベルと見なし，いずれかの言語現象を一つ以上含む 834 文（全体の 44%）に対して代表ラベル種別の頻度を算出した．図 1 に示した通り，UGC には固有名詞やスラングのように語彙の分布に影響する現象が特に多く見られ，次いで表記の揺らぎに由来する現象が見られた．この結果を受け，以降のすべての実験は高頻度な 4 現象（固有名詞・名詞の省略・かな表記・非正規形）を対象に分析を行った [*6]．

次に，各現象が NMT システムに与える影響を調査するため，元文および現象置換後文の 2 種類を入力としたときのシステムの性能差の比較を行った．表 3 に比較の結果を示す．

固有名詞・かな表記を含む文に対しては，それらの現象を置き換えたデータセットで評価した場合，すべてのモデルにおいて BLEU が向上した [*7]．ただし，学習データの制約がない unconstrained モデルにおいて，固有名詞データセットでは置き換えを行った際の BLEU の向上幅が大幅に小さくなった（+4.2 → +1.4）のに対し，かな表記では変動が見られなかった．このことは，かな表記の翻訳精度が学習データ量に依拠せず，特別な対処を必要とする現象であることを示唆している．しかし，固有名詞に関しても，学習データを拡充しさえすれば NMT システムへの影響を緩和できると結論付けるのはやや早計である（詳しくは後述の 6. 節で議論する）．また，名詞の省略や非正規形データセットにおいては，置き換えによる BLEU の差分が比較的小さい結果となった．前述のように，非正規形とされた表現はかな表記等を包含する場合が多く見られ，置き換え後に他現象に起因する精度低下が生じた可能性がある．

表 3: 現象毎評価データセットの BLEU

			constrained		unconstrained	
	size		base	tuned	base	tuned
固有名詞	290	orig.	12.7	14.6	18.1	17.1
		repl.	16.9	18.5	19.5	18.3
		diff	+4.2	+3.9	+1.4	+1.2
名詞の省略	97	orig.	11.2	12.2	11.4	14.0
		repl.	12.0	12.7	13.1	13.3
		diff	+0.8	+0.5	+1.7	-0.7
かな表記	77	orig.	11.8	12.5	13.6	12.0
		repl.	13.4	14.9	15.4	14.8
		diff	+1.6	+2.4	+1.8	+2.8
非正規形	72	orig.	11.6	12.2	10.8	11.1
		repl.	10.3	12.9	11.7	12.2
		diff	-1.3	+0.7	+0.9	+1.1

5.2 定性分析

各現象について，4 モデルのうち最も BLEU が高かったモデルの出力を対象に，表現の置き換えによる出力の変化を分析した．表 4（a）はかな表記が見られた文と該当箇所を訂正した文に対するシステム出力の一例である．この例では，モデルが，ひらがな表記された「かしこい」を “crazy” と誤訳してしまったのに対し，置換後の漢字表記された「賢い」は適切に翻訳することができた．このように，多くの文脈に出現することで多義性を伴うと思われるひらがな表記を，曖昧性の少ない漢字表記へ正規化することで正しく訳出可能になる例はデータセット中に多く見られた．表 4（b）の例は名詞の置き換えにより出力が改善した例である．多くの熟語を構成し得る複数の漢字の組みあわせ（「漢」＋「検」）から，その構成熟語が明らかになり曖昧性が解消されたことで適した訳を出力できていた．省略形が英頭文字の例では，置換後の表現が一般的に用いられるものでない場合も見られた（表 4（c）中 “NPC” → “nonplayer characters” など）．これらの中には，置換後（非省略形）の訳も意味的に十分であるにも関わらず，参照訳中では省略形表記（NPC）となっていたために，BLEU による評価が低スコアとなる問題が生じるものが複数見られた．このことから，名詞の省略に対する対訳の許容可能性を測定する上では，単一参照訳 BLEU による評価では不十分だと考えられる．

*6　スラング・スタイル変化については，訂正の恣意性が極めて高く，一致を取りづらい現象であることから本評価実験の対象外とした．

*7　固有名詞データセットでは，参照訳の語長変化により BLEU の比較が公平でない可能性がある．しかし，今回構築したデータセットでは置換による参照訳語長の変化は+1.03 語/文であり，置換前との語数の変動が小さいことから，報告した精度差は “a certain” の n-gram 一致による影響以上のものだと考えている．

表 4: システム出力例

(a) かな表記	
Ja	編集部は { かしこい/**賢い** }
orig. output	The editorial department is crazy.
repl. output	The editorial department is **smart**.
(b) 名詞の省略	
Ja	自分はまだ { 漢検/**漢字検定** } を許してない
orig. output	I still haven't forgiven the Chinese prosecutor's Office.
repl. output	I still haven't forgiven the **Kanji test**.
(c) 名詞の省略	
Ja	ってのはよくいわれることだけど、実際は {NPC/**ノンプレイヤーキャラクター** } がいないのがもっとつらい
orig. output	That's what people talk about, but actually it's more painful that there is no NPC.
repl. output	That's what people talk about, but actually it's hard to have no **nonplayer characters**.

6. 考察

6.1 固有名詞の時系列的検証の必要性

固有名詞データセット中に出現する固有名詞について，学習データ中に出現する割合（カバー率）を算出した．対象とした 278 の固有名詞のカバー率は，constrained データセットにおいて 58.3%，unconstrained データセットでは 75.2%であった．学習データの増加に伴い，学習時に既出の事例が増加したことは評価データ間のスコア差縮小の一因だと考えられる．一例として,「カルロス・ゴーン」という表現は JParacrawl のみに出現しており，学習データに JParacrawl を含まない constrained モデルにおいては誤ったスペルで転写されてしまったのに対し，unconstrained モデルでは正しく出力できていた．しかし，対象表現の出現文脈から，カバー率向上分の一部は時系列的要因に帰着する可能性が示唆される．Michel ら [2] は MTNT データセットの作成過程において，2017 年 11 月から 2018 年 3 月のコメントを用いたと報告している．一方，次の JParacrawl 中の文はその文脈から 2018 年 12 月以降の投稿と推測される．

「*2018* 年 *12* 月日本の出来事—*nippon.com (...)* カルロス・ゴーン前日産自動車会長の特別背任容疑での再逮捕、政府の国際捕鯨委員会 *(IWC)* 脱退表明など、*2018* 年 *12* 月の日本の出来事を振り返る。」

このような，評価データ以降に作成された事例の存在は，NMT システムの固有名詞による影響を過小評価する一因となり得る．翻訳システムは実用上，対訳コーパスの入手時には未知であった新たな固有名詞にも対処する必要がある．固有名詞の翻訳システムへの影響をより適切に評価するには，学習データおよび評価データの時系列的な分離が必要だと考える．

6.2 かな表記の多義性

かな表記データセット中の該当表現についても，同様に学習データ中のカバー率を調査した．置換後の表現のカバー率は constrained，unconstrained の 2 種類の学習データにおいていずれも 91.4%であった．一方で，置換前の「かな」で表記された表現ではカバー率に 6%程度の差が見られた．しかし，固有名詞とは異なり，評価対象データのより多くの事例を学習している unconstrained モデルにおいても，評価データ間のスコア差に改善は見られなかった（表 3）．これは，表音文字である「かな」が多様な文脈で出現することにより高い曖昧性を有することが原因だと考える．かな表記は日本語に特有の現象であるものの，類似の問題は他言語においても見受けられる．特に,「**きゅうり**ょう」（給料，きゅうり）など，表記の揺れ等により一部またはその全体が他の実在単語の一般的表現となっている場合，適切な翻訳をどう得るかは今後の課題である．

7. おわりに

本研究では，UGC の高品質な翻訳の実現に向け，様々な言語現象が NMT システムに与える影響について初めて詳細な分析を行った．具体的には，言語現象の存在を打ち消したデータセットを構築し，元の評価データとの精度比較を通して，現象に注目した翻訳システム評価を行った．実験の結果,「かな表記」の翻訳精度は学習データ量に依拠せず，特別な対処を必要とする現象であることを明らかにした．また，今回構築した評価データセットの公開を行った [*8]．今後の方針としては，頑健性評価におけるデファクト化に向けた各現象データセットの拡張，各現象に対する頑健性向上手法の検討があげられる．

謝辞

本研究の一部は，JSPS 科研費 JP19H04162 の助成を受けたものです．

参考文献

[1] Thang Luong, et al. Effective Approaches to Attention-based Neural Machine Translation. In *EMNLP 2015*, pp. 1412–1421, 2015.

[2] Paul Michel and Graham Neubig. MTNT: A Testbed for Machine Translation of Noisy Text. In *EMNLP 2018*, pp. 543–553, 2018.

[3] Huda Khayrallah and Philipp Koehn. On the Impact of Various Types of Noise on Neural Machine Translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 74–83, 2018.

[4] Philipp Koehn, et al. Findings of the WMT 2018 Shared Task on Parallel Corpus Filtering. In *WMT 2018*, pp. 726–739, 2018.

[5] Xian Li, et al. Findings of the first shared task on machine translation robustness. In *WMT 2019*, pp. 91–102, 2019.

[6] Antonios Anastasopoulos, et al. Neural Machine Translation of Text from Non-Native Speakers. In *NAACL 2019*, pp. 3070–3080, 2019.

[7] Myle Ott, et al. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *NAACL 2019*, pp. 48–53, 2019.

[8] Ashish Vaswani, et al. Attention is All you Need. In *NIPS 2017*, pp. 5998–6008, 2017.

[9] Soichiro Murakami, et al. NTT's machine translation systems for WMT19 robustness task. In *WMT 2019*, pp. 544–551, 2019.

[10] Makoto Morishita, et al. JParaCrawl: A large scale web-based japanese-english parallel corpus. 2019.

[11] Rico Sennrich, et al. Neural machine translation of rare words with subword units. In *ACL 2016*, pp. 1715–1725, 2016.

[12] Kishore Papineni, et al. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL 2002*, pp. 311–318, 2002.

*8 `https://github.com/cl-tohoku/type_wise_robustness_eval`