一般論文

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[†]・三田　雅人[††,†]・阿部香央莉[†]・塙　一晃[††,†]・
森下　睦[†††,†]・鈴木　潤[†,††]・乾　健太郎[†,††]

ニューラル機械翻訳 (NMT) の登場により，ニュース記事など文体の整った入力に対する翻訳の品質は著しく向上してきた．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 NMT の翻訳には依然として多くの課題が残されている．異文化・多言語交流の促進に向けた機械翻訳システムの活用には，そうした特異な入力を正確に扱うことのできる翻訳モデルの構築が不可欠である．近年では，UGC における翻訳品質の向上に向けたコンペティションが開催されるなどその重要性は広く認知されている．一方で，UGC に起因するどのような要因が機械翻訳システムの出力に悪影響を及ぼすのかは明らかでなく，偏在するユーザコンテンツの翻訳に向けた確かな方向性は依然として定まっていない．そこで本研究では，言語現象に着目した日英機械翻訳システムの頑健性測定データセット **PheMT** を提案する．特定の言語現象を含む文に特化したデータセットにより，当該表現の翻訳正解率，および正規化に基づく翻訳品質の差分を用いた精緻なエラー分析を可能にする．構築したデータセットを用いた評価により，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 モデルにおいても十分に扱えない，対処すべき言語現象の存在を明らかにする．

キーワード：機械翻訳，ソーシャルメディア，頑健性，データセット

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

Ryo Fujii[†], Masato Mita[††,†], Kaori Abe[†], Kazuaki Hanawa[††,†], Makoto Morishita[†††,†], Jun Suzuki[†,††] and 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 make better use of NMT for cross-cultural communication, one of the most promising directions is to develop a translation model that correctly handles these informal expressions.

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

Though its importance has been recognized, it is still not clear as to what creates the large performance gap between the translation of clean input and that of UGC. To answer the question, we present a new dataset, **PheMT**, for evaluating robustness of MT systems against specific linguistic phenomena in Japanese-English translation. We provide more fine-grained error analysis about the behavior of the models with the accuracy and relative drop in translation quality on the contrastive dataset specifically designed for each phenomenon. 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.

Key Words: *Machine Translation, Social Media, Robustness, Dataset*

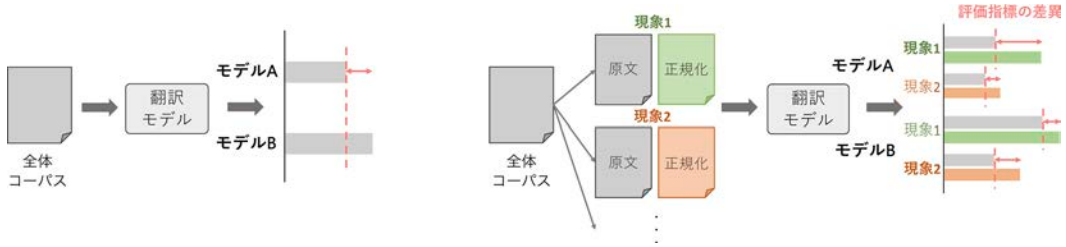
1　はじめに

ニューラル機械翻訳 (Neural Machine Translation, NMT) の発展 (Luong et al. 2015; Vaswani et al. 2017) により，ニュース記事のような文体の整った入力に対する翻訳品質は著しく向上し，一部の言語対においては既に人間の翻訳に匹敵するレベルにまで到達したと言われている (Hassan et al. 2018; Barrault et al. 2019)．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に見られるようなユーザ生成コンテンツ (User-Generated Contents, UGC) に対する NMT の適用可能性は依然として極めて限られている (Michel and Neubig 2018)．

一方で，ソーシャルメディアなどの普及に伴い，UGC が我々の日常生活に与える影響は非常に大きなものとなっている．例えば，それらのサービスへの投稿はユーザの購買行動の決定にも重大な影響を及ぼすことが報告されている[1]．そのような背景において，Berard et al. (2019a) は Foursquare[2]に寄稿されたレストランのレビューに着目し，異文化交流の促進に向けて実応用を見据えた翻訳タスクを設計した．

近年では UGC に対して頑健な翻訳システム構築への関心の高まりと共に，ソーシャルメディア上のテキストの翻訳精度を競うコンペティションも開催されている (Li et al. 2019)．当該コンペティションにおいて，翻訳の評価は従来の機械翻訳出力に対する評価と同様に，あるひとつの包括的なデータセットに対してひとつの全体スコアを付与する手法で行われている（図 1a)．しかし，我々はこの評価における改善は必ずしもモデルの頑健性を説明していないと考える．例えば，訓練データ規模の異なる 2 つのモデル出力に対して BLEU スコアの比較を行い，スコアの改善をもってあらゆる事象に対する頑健性を結論付けることは危険である．UGC における機械翻訳システムの性能向上への端緒を見出すためには，翻訳品質の低下を招く要因を明らかにし，実際にそれらが改善されていることを示すことのできる確かな基盤が必要である．

[1] https://stackla.com/resources/reports/the-consumer-content-report-influence-in-the-digital-age/
[2] https://foursquare.com/



(a) 一般的な評価　　(b) 現象毎データセットによる評価

図 1　一般的な機械翻訳評価の手法と我々のデータセットによる現象毎評価の比較.

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，日本語ユーザコンテンツに含まれる特定の言語現象が出力に及ぼす影響を調査する．具体的には，既存の評価データセットに対して，新たに固有名詞，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アノテーションを付与することで，言語現象毎評価データセット **PheMT** (**Phe**nomenon-wise Dataset for **M**achine **T**ranslation Robustness) を構築した（図 1b）．文中の特定の表現に注目することにより，ある言語現象を正しく扱えるかどうかを翻訳正解率を用いてより直接的に測定することを可能にした．また，ある現象がモデルに及ぼす影響は，その現象の存在を取り除いた場合との差分により評価できるというアイデアのもと，人手による当該表現の正規化を行った．原文と正規化後の文を入力した際の評価指標の差異の測定により，個々のモデルが元来有する表現能力の違いに起因する影響を取り除き，当該表現が翻訳文全体の品質に与える影響を測定する．構築したデータセットを用いた評価と分析を通して，UGC という限定されたドメインにおいても依然として十分に翻訳することのできない対処すべき現象が存在することを明らかにする．

本論文の貢献は以下の 2 点である．

(1) 日英機械翻訳における詳細なエラー分析のための第一歩として，日本語ユーザコンテンツに頻繁に現れる言語現象に着目した評価データセットを構築した．

(2) 構築したデータセットを用いた評価と分析により，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 モデルにおいても，依然として多くの課題が残されていることを明らかにし，今後の機械翻訳評価における一つの方向性として言語現象毎評価の可能性を示した．

2　関連研究

機械翻訳分野において，入力文中のノイズに頑健なシステムの構築に対する関心が高まっている．これに対して，Michel and Neubig (2018) は UGC の代表例となるオンラインディスカッションサイト Reddit[3]に投稿されたコメントに対して，プロの翻訳家による参照訳を付与することで英日翻訳を対象とした MTNT (**M**achine **T**ranslation of **N**oisy **T**ext) データセットを構築した．彼らは，MTNT データセットの原言語文が既存の他の機械翻訳ベンチマークと比較して誤植（タイポ）や文法誤りなどのノイズを多く含むことを明らかにすると共に，ベースラインモデルによる翻訳実験の結果から UGC に対する翻訳の難しさを示した．また，このデータセットは近年開催された機械翻訳システムの頑健性に関するコンペティションにおいても，訓練および評価のために用いられた[4]．

しかし，MTNT データセットには，ニュースの見出しのように文語調なものから，隠語を含む特定コミュニティ内の崩れた会話まで多種多様な文体やドメインが含まれている．また，付加された翻訳文においても，その解釈の難しさから未訳や誤訳を含むものが一定数存在し，品質は玉石混淆であると言える．このようなデータセットを評価基盤として，機械翻訳における代表的な評価指標である BLEU スコア (Papineni et al. 2002) を用いた比較を行った場合，スコアの改善が実際にどれほどノイズに対する頑健性に起因するのかを推定することは難しい．事実，コンペティションにおいて最も良いスコアを獲得したチームである Berard et al. (2019b) は，UGC に頻出する大文字・小文字の混同に対処するため各単語の先頭にタグを付与するなどの工夫を行い，人手評価をもって入力に対する頑健性が向上することを示したが，BLEU スコアを用いた評価では，単に訓練データ中からペアとして不十分なものを取り除くコーパスフィルタリングが最も改善に寄与したと報告している．コーパスフィルタリングは，データ駆動の機械翻訳システムにおいては不可欠な技術であるものの (Koehn et al. 2018; Junczys-Dowmunt 2018)，これはパラレルコーパスの構築にあたって生じたノイズへの対処を目的としており，本質的に**入力に対する頑健性**を対象としてはいない．したがって，現状のデータセットおよび評価をもって UGC という特異な入力に対するシステムの優位性を帰結することは早計であり，真に頑健な機械翻訳システムの構築に向けては，低コストで再現性が高く継続的な改善を実現可能な評価基盤を築くことの重要性は高いと考える．これらの問題意識のもと，本研究では入力に見られる特殊な言語現象に対する改善を重点的に測定できるような評価用データセットを構築することを目的とする．

言語現象という観点から誤訳の原因究明を目指すアプローチは，誤植 (Heigold et al. 2018; Belinkov and Bisk 2018; Karpukhin et al. 2019; Niu et al. 2020)，文法誤り (Nagase et al. 2011;

[3] https://www.reddit.com
[4] http://www.statmt.org/wmt19/robustness.html

Sennrich 2017; Anastasopoulos et al. 2019) からジェンダーバイアスの存在 (Stanovsky et al. 2019) に至るまで広く受け入れられている．例えば，Isabelle et al. (2017) は英語からフランス語への翻訳において，主述の一致など構造の差異から生じる 20 種類以上の細分類を言語現象的観点から設定し，当該箇所の翻訳精度によりモデルを評価した．モデルの汎化性能の測定においては様々な性質のデータを用いて評価することが一般的であり，それぞれが単一の言語現象を対象とした少数データセットによる分析，という彼らの提案は機械翻訳システムに対する理解を促進する上で新たな側面を提供したと言える．一方で，彼らの手法は人手による絶対評価を要するため，言語学に関する高度な知識を持った公平な評価者の選定や評価データの拡張において課題がある．

この問題に対し，Sennrich (2017) は対照データセットによる相対比較を用いた自動評価のアプローチを提案した．この研究では，それぞれの入力文に対し，正解の参照訳とその一部を改変した対照的な参照訳を作成し，モデルが正解に対して高い生成確率を与えることができた文の割合をもって評価が行われた．後に，Bawden et al. (2018) は同様の手法を用いて，先行文脈を利用した機械翻訳における照応代名詞の影響を検証した．しかし，彼らが指摘するように，この手法はモデルが 2 つの参照訳を正しく順位付けできた場合にも，最も尤度の高い生成文が誤りを含まないことは保証しない．さらに，これらの手法は入力に曖昧性のない整った文を想定しているため，そのような特徴を有しない UGC の分析に適用することは難しいという問題がある．

一方で，対照的なのは必ずしも参照訳である必要はない．Heigold et al. (2018) は，誤植を模した人工的な誤りをルールにより作成し入力文に加えることで，**入力文中の**誤りの存在がモデルの翻訳品質の低下にどれだけ寄与するかを調査した．さらに，Belinkov and Bisk (2018), Karpukhin et al. (2019) はウェブサイトの編集履歴を用いることで，人間が生成しやすい誤りの分布を再現し，検証の範囲を自然な誤りに拡張した．しかし，これらの研究では誤りは疑似的なものであり，実際に人間が記述したテキストに対するモデルの頑健性は検証されていない．これに対して，Anastasopoulos et al. (2019) は文法誤りに対するベンチマークである JFLEG コーパス (Napoles et al. 2017) に翻訳文を付与することで，人間が生成した誤りを直接入力文として用いた．彼らは，テスト時に含まれる誤りと同種の誤りを含む訓練事例を用いてモデルを学習することで，他種の誤りを加えて学習したモデルに比べ高い精度を達成することを示した．

このように，機械翻訳システムの頑健性に関して特定の観点に着目した研究は多く行われている．しかし，これらの工夫をもってしてもなお，我々が取り組む UGC の翻訳においては，ニュース記事などを翻訳する場合に比べて著しい品質の低下が見られる．この背景には，UGC に特有の難しさが体系的に分類されていないことや，UGC を模したデータへの変換が極めて難しいことがあげられる．これに対して，我々は UGC に頻繁に見られる言語現象の定義を行い，UGC 上の文からそれらの現象を取り除く正規化を行う事で対照データセットを作成する．

また，先行研究の多くは文字体系を共有する比較的近い言語対を対象としており，我々の知る限りでは日本語-英語のような類似しない言語対における詳細なエラー分析のための基盤は未だ構築されていない．しかし，固有名詞の翻字に示されるように，文字体系を共有しない言語対に特有の問題も生じるため，取り扱う言語対に応じた柔軟な拡張が必要であると考えられる．我々のデータセットが，この極めて難しい言語対の翻訳に対する新たなアプローチの一助となることを期待する．

3　現象毎データセットの構築

本節では現象毎データセット構築の手順について説明する．我々は，UGC における機械翻訳システム評価のための既存のベンチマークである MTNT データセットに新たに言語現象のアノテーションを行うことでデータセットを構築した．図 2 に一連の手順の概要を示す．データセットの構築は主に，品質担保のための事前データ選定と現象該当箇所の抽出および正規化の 2 つの段階から成る．まず 3.1 節では，翻訳品質の基準となる十分性スコアの定義と，MTNT データセットに対する人手のアノテーションを行った結果として得られたスコアの分布について考察する．後述する 3.2 節では，評価の対象とする言語現象の定義および，表現の抽出と正規化の手順について実例を交えて展開する．

3.1　品質担保のための事前データ選定

現象毎データセットを構築するにあたっては，現象を含む文を収集しそこに新たに翻訳を付与する方法と，既存のパラレルデータに現象のアノテーションを行う方法の 2 つが考えられる．

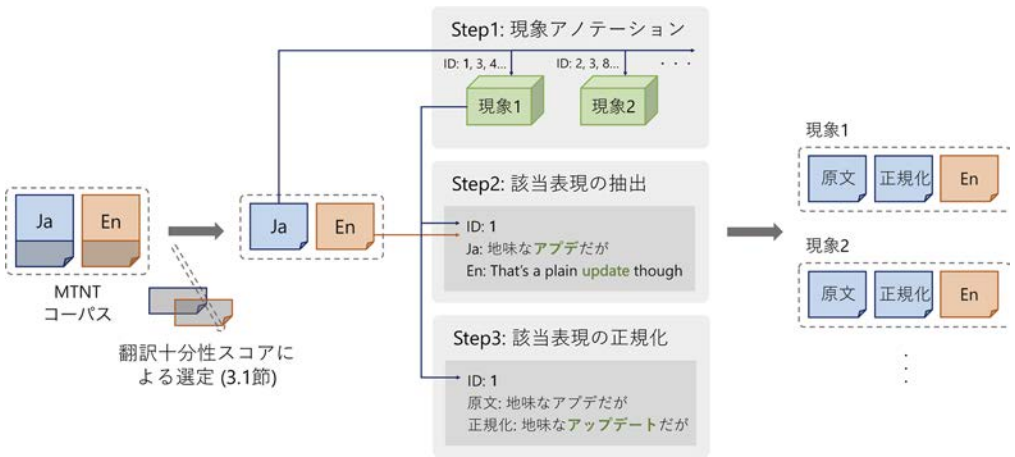


図 2　現象毎データセット構築の流れ．

前者の方法には，原言語に含まれるニュアンスや文体を目的言語にも保持するかという観点から参照訳をコントロール可能であるという利点がある．例えば，「かわいいいい」のような表現では cute，cuuuute のいずれもが正解となり得るため，それぞれの形式にどの程度対応可能であるかを個別に把握することの重要性は高い．しかし，UGC に対して高品質な翻訳を大規模に付与する方法は自明ではなく，このようにボトムアップにデータセットを構築することは高いコストを伴う．そのため，我々は既存のベンチマークである MTNT データセットにラベルを付与し，分類・正規化を行うことでデータセットの構築を試みた．ここで，MTNT データセットでは各言語対に対して翻訳方向毎にデータセットが用意されている．このことから，我々は対訳にはニュアンスの保持を期待しないものとしてデータセットの構築・分析を行った．また，MTNT データセットを複数の現象に細分類する場合，評価用データの文数は限られているため，個々の現象データセットが極めて少ない文数に分割されてしまう恐れがある．そこで，我々は MTNT データセットの訓練および開発用データを含めた全データを，評価用データ (PheMT) として用いることにした．しかし，通常，訓練データは規模が大きいため人手による品質の評価を経ていない場合が多く，実際に人手で数件確認した限りでも，MTNT データセットの訓練データには多くの未訳や誤訳が含まれていた．そこで，我々は翻訳品質の基準として十分性スコアを定義し，MTNT データセット中の全文に対し，クラウドソーシングを用いて評価を行うことで現象毎データセットの品質を担保することにした．

十分性スコアの付与に先立って，まず簡単なルールベースのフィルタリングを行った．具体的には，最低限の品質を満たさない文対として，(i) 対話コーパスのフィルタリングのために定義された不適切語リスト[5]中の単語を含む (ii) 原言語文と目的言語文が一致する（未翻訳） (iii) 他の文対と完全に一致する のいずれかを満たすものを除外した．さらに，アノテーションタスクの平易化のため (iv) 一方もしくは両方の言語が 1 語のみ，もしくは 80 語以上からなるものを除外した．フィルタリングの適用後，残った文対について 1（不適切）– 5 （適切かつ流暢）の 5 段階のリッカート尺度で分類するタスクを行った．

各スコアへの分類基準の策定にあたっては，機械翻訳システムの出力を人手評価する際に一般的に用いられる，適切性 (adequacy) と流暢性 (fluency) の 2 つの観点 (White et al. 1994; Li et al. 2019) を用いた．しかし，我々が今回評価対象とする翻訳文は人手で生成されたものであるため，流暢性に関してはある程度担保されていると考えるのが自然である．そこで，流暢性の観点として翻訳語（機械翻訳らしさ）を参照し (Graham et al. 2019; Freitag et al. 2020)，表 1 に示す 5 段階の基準を設定した．

タスクの実行に際し，翻訳文の人手評価は本来目的言語の母語話者を評価者として行われるべきであるが，UGC の文意を正確に把握するためには当該言語に対する高度な理解を必要とす

[5] https://github.com/1never/open2ch-dialogue-corpus

るため，英語が堪能な 10 名の日本語母語話者に評価を依頼した[6]．評価者によるスコアのぶれを低減するため，1 文あたり 3 名の評価者を割り当て，平均をもって各文の十分性スコアとした．

図 3 に本タスクによって付与された MTNT データセットの各分割に対する十分性スコアの分布を示す．この結果から，青色および黄色で示される学習，開発用セットにおけるスコアの分布が，評価用に作成された他の分割（テスト，ブラインドセット[7]）に比べ，低スコア側で高くなったことがわかる．このことは，評価データを単純に拡張する事が招く品質の低下と，フィルタリングの必要性を示唆していると言える．現象毎データセットの構築にあたっては，両言語により伝わる情報を等価と見なすことのできる，平均スコアが 4.0 以上の文対のみを用いることとした．表 2 には十分性スコアの各スコア帯における原文と翻訳文の一例を示す．

また，我々はこれらのアノテーションが妥当かつ信頼性の高いものであるかを検証するため，アノテーションの一致率を調査するとともに，一部の文対に対して英語母語話者により同様の品質評価を実施した．まず，アノテータ間一致率について，順序尺度に最適化されていることや，欠損値を含むデータにも適用可能なことから Krippendorff の alpha 係数 (Krippendorff 2011) を

スコア	定義
5	原文（日本語文）の意味を完全に伝える流暢な翻訳
4	訳抜けや誤訳を伴わないが，過度に逐語的な翻訳
3	局所的な未訳や誤訳を含むが，意味の概観を捉えることが可能で許容できる翻訳
2	フレーズや文レベルの重大な未訳または誤訳を含む，あるいは誤った解釈に基づく翻訳
1	原文（日本語文）の意味を全く表さない誤訳

表 1　翻訳十分性スコアの評価基準．

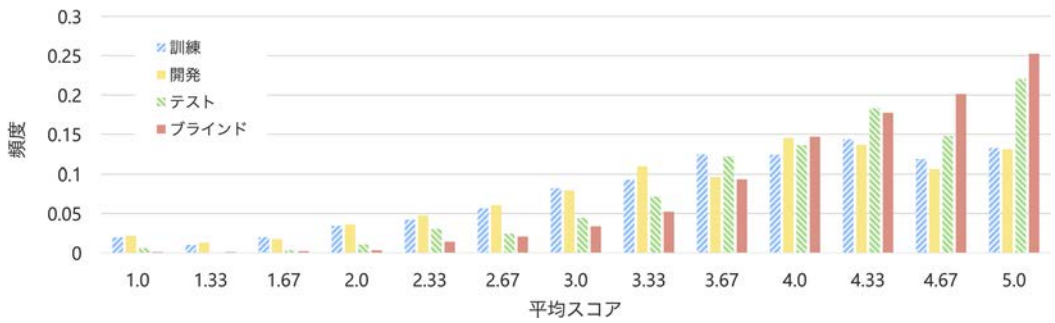


図 3　MTNT データセットの各分割に対する十分性スコアの分布．

[6] 1 名あたりの担当文約 2,200 文に対し，標準報酬を 2 万円に設定した．10 名の募集に対し，80 名以上の応募者の中から，職業としての翻訳経験，海外在留経験，もしくは同等のスキルを有するワーカを選出した．

[7] コンペティションにおいてシステムの順位を決定するために用いられたデータセット．評価期間の終了まで参加者には公開とされた．

平均	スコア	原言語文	目的言語（翻訳）文
5.0	5, 5, 5	割と気に入って四年くらいそのままだわ。	I somewhat like it so I've left it that way for 4 years.
4.0	4, 4, 4	今って、ひと昔のオタクっぽい人ってどこに行けばいいの？	Now, where can people who are like otaku from just a bit ago go?
3.0	2, 2, 5	部屋着なに着てる？	What do you wear to dress up?
2.0	3, 2, 1	カレー屋ネパール人「二人前頼めや」	A two person reservation.
1.0	1, 1, 1	仮想通貨とかいうオワコン・・・とでも言うと思ったか？	I made 100 thousand into 400 thousand today.

表 2　十分性スコアの各スコア帯における翻訳文例.

測定した結果，一致率の値は 0.32 となった．これは，Landis and Koch (1977) の基準において fair agreement （まずまずの一致）とされる値であり，一致率としては少し低いと考えられる．原因としては，評価対象文である MTNT データセットの翻訳が人手で生成されていることから意味の等価性に重点を置いた評価基準を設定したものの，実際には機械翻訳にかけたような流暢でない翻訳文も散見され，文法性の許容範囲に主観の入る余地が生まれてしまったことが想定される．さらに，各ワーカの付与したスコアの分布に着目すると，10 名のワーカの平均点の範囲は 3.32–4.44 点と広いものの，四分位範囲は 0.46 点に収まっていた．これは，図 4 に示すように評価対象文の大多数に高い点数を付与するようなワーカの存在に起因すると考えられる．そこで，スコアの平均が最も高い・低い各 2 名の結果を外れ値として除外し，再度 alpha 係数を測定したところ，値は moderate agreement （中等度の一致）となる 0.48 まで向上した．ワーカの質には依然として議論の余地が残るものの，利用可能な文数とのトレードオフにより信頼性の高い部分集合を用いるなど品質の調整は可能である．

次に，妥当性に関して，具体的にはクラウドワーカによる平均点が 4.0 以上・未満の文からそれぞれ 50 文対を無作為に取り出した合計 100 文対に対して，翻訳業を生業とする英語母語話者にクラウドソーシング実施時と同様の指針を提示して評価を依頼した．評価の結果を図 5 に示す．図から，平均点が 4.0 以上であった 50 文対に対して新たに 3 点以下が付与されたものは 7 文対であり，4.0 未満の 28 文対に対して 4 分の 1 と少ないことから評価の妥当性は保たれていると考えられる．また，表 3 にはとりわけ点数の低かった 3 文対とその判断の根拠（原文）を示す．この結果から，スコアに乖離が生じた例の中には文脈情報の不足により評価が困難なものも多いように思われる．文章単位の情報が利用可能な状況では改善の余地はあるものの，そのような例を完全に取り除くことは難しいと考えられる．

以上の結果を踏まえ，言語現象の評価に用いる文を可能な限り多く確保するため，今回の実験ではアノテーションの信頼性によるフィルタリングは行わなかった．その結果，条件を満たした文数は評価対象とした 7273 文の約 57.1%にあたる 4152 文であった．また，訓練データを評価のために用いる際のもう一つの問題点として，複数の翻訳文を持つ原言語文の存在があげられる．特に，MTNT データセットはその構築過程において，一連の発言であるコメント単位で翻訳を行い，文単位に自動で分割する処理を経るため，アラインメントの誤りに起因する重

図 4　各ワーカの付与した十分性スコアの分布.

複部分を多く含む特徴がある．そこで，我々は原言語文を空白で区切った際の最初の分割が同一のものを集計し，最も高いスコアが付与された一文のみを参照訳として保持することで対処した．これにより，現象ラベルの付与対象は 3896 文となった.

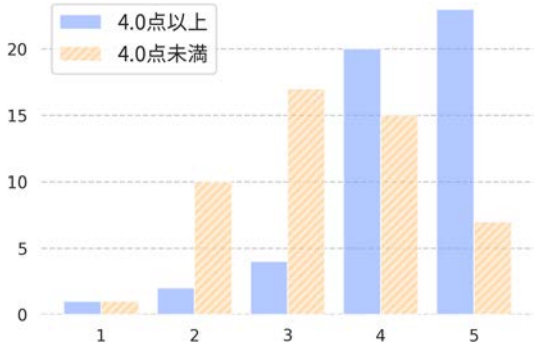


図 5　平均点 4.0 以上・未満の各 50 文対に対する英語母語話者による品質評価.

スコア	原言語文	目的言語（翻訳）文	判断の根拠
2	コインチェックが全ての通貨の出金を停止	Coincheck stops all currency financing.	This should probably say ” ”withdrawals” ” rather than ” ”financing” ”, I think it means that transactions to withdraw cryptocurrency as normal currency have been stopped, but it’s hard to infer that meaning from the translation
2	ござるは邪道なりよ	”Gozaru” leads to bad things.	If this means something like ” ”Speaking like a samurai and using ’gozaru’ all the time in sentences will lead you down a bad path” ” then it would be difficult to translate in English anyway, but without extra context, this would be hard to understand for someone who doesn’t speak Japanese
1	3LDK = LLL	3LDK=LLL	I mean, this technically contains all the information from the original sentence, which is just alphanumeric characters, so one could hope for 100% equivalency, but we don’t refer to properties in this format in English, so it wouldn’t make sense unless you localised it to be ” ”a 3-bedroom property” ”.

表 3　低スコアが付与された文対とその判断の根拠.

3.2　評価データセットの構築手法

(i) 言語現象ラベルの定義　評価の対象とする言語現象のラベルを定義するにあたって，我々は UGC に特有のどのような要因が他の言語処理タスクにおいて問題となっているかを調査した．Sasano et al. (2013) は日本語形態素解析の文脈において，インターネット上のテキストに含まれる標準的でない綴りの影響に着目している．彼らは，母音の長音化や小文字化などのルールを人手で設計し，ルールに基づいた正規化適用後の表現が事前定義済みの語彙に含まれる場合に，形態素ラティスに追加することで解析誤りを抑制できることを報告した．さらに，同様のルールに基づいて作成した正規-崩れ文字列パターンのシードデータからアラインメントを学習することで，より複雑な崩れ表記にも対応できることが示されている (斉藤 他 2014)．また，ニューラルネットワークを利用したテキスト正規化モデルにおいても，疑似データの作成時に類似のルールを適用することで精度の向上が確認されている (Ikeda et al. 2016)．しかし，これらの崩れ表記が機械翻訳などの言語横断的タスクに及ぼす影響は未だ十分に検討されていない．そこで，我々はこれらの先行研究に倣い，口語表現および異表記の 2 種類の現象ラベルを定義した．

さらに，我々は UGC を含む様々なテキストにおいて頻繁に見られる固有名詞および名詞の省略の 2 種類の現象に着目した．固有表現が機械翻訳システムに及ぼす影響については翻字や転写の文脈で注目を浴びつつある (Shao and Nivre 2016; Rosca and Breuel 2016; Ugawa et al.

2018)．我々は，これらの現象が実際に UGC においてどれだけ現れるかを検証するため，予備実験として MTNT データセットの訓練データから無作為に取り出した 500 文に対するアノテーションを行った．実験の結果から，名詞の省略について全体の 10%以上，固有名詞については 40%以上の文に含まれる極めて頻出の現象であることが分かった．

我々は，以上の 4 種類の現象について最新の機械翻訳システムに及ぼす影響を調査する．各ラベルの具体的な定義とその分類例を表 4 に示す．

(ii) 該当表現の抽出と正規化　言語現象毎データセットの構築にはクラウドソーシングを用いた．問題の難易度およびワーカ間の一致率を考慮し，アノテーション手順を各文への現象ラベルの付与，該当表現およびアラインメントの抽出，正規化の 3 つの段階に分けてデータセットを構築した．また，回答の品質を担保するため，すべてのタスクにはチェック設問を設け，各設問に対しチェック設問を通過した 5 名のワーカの回答を集約した．

初めに，3.1 節の十分性スコアにより選定した MTNT データセット中の文対に対して，原言語（日本語）文を表 4 に定義する 4 種類の現象ラベルに分類するタスクを行った（図 2 Step1）．各設問は，ある文に対して各現象の有無をチェックする 4 つの小設問からなる．ワーカの過半数にあたる 3 名以上が同一の現象に対して「含む」と回答した場合のみ，その現象を含むと見なした．

続いて，現象ラベルとその該当表現の対応づけを行った（図 2 Step2）．具体的には，日本語文と現象ラベルのペア 1 つに対し，最大 5 箇所を本文中の形式で抜き出すタスクとして設計した．つまり，前述の分類タスクにおいてある文が複数の現象を含むと分類された場合，注目する現象毎に別の設問として取り扱われる．また，抽出された各表現について目的言語中から対応する表現（アラインメント）の抜き出しを行った．これらのタスクに対する回答は過半数の完全一致が得られたもののみを使用した．

言語現象ラベル	定義	分類例
固有名詞	人名，作品名，企業・組織名，ブランド名，国・地域名など，同一物が複数存在しないもの	安倍首相，アナと雪の女王
名詞の省略	正式名称の一部を短縮することにより作られた語，頭文字を用いた表記	アプデ，WHO
口語表現	“っ”（促音），“ん”（撥音），“ー”（長音）または母音の挿入・置換・脱落，あるいは発音の崩れにより辞書的な表記から派生した語	ねむーーい，かなちい
異表記	ひらがなやカタカナで書かれた表現のうち異なる表記（漢字など）が一般的であるもの，小文字化された表現	アリガトウ，ぃぃよ

表 4　言語現象ラベルの定義と分類例．

さらに，各現象の影響を取り除き対になる評価文を作成するため，抽出された表現の正規化を行った（図 2 Step3）．ここで「正規化」とは，表 4 の定義に従い辞書中の表記から逸脱した表現を，対応する標準形の表現に書き直すことを意味する．つまり，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逆向きの変形を適用する．例えば，名詞の省略に分類される「アプデ」という表現は，この表現を名詞の省略ではなくするような変形，つまり省略元の表現への書き換えにより「アップデート」に正規化される．同様に，長音記号の挿入により口語表現となる「ねむーーい」という表現は，長音記号を除去することで「ねむい」に正規化される．ここで，口語表現と異表記は標準形と同音であるか否かにより定義上排他的であるものの，一部の口語表現では正規化適用後により一般的な漢字表記（上記例「眠い」）を持つ異表記に転化する場合も見られた．このような例では，口語表現の範囲外となる漢字化は行わないよう指示し，当該現象の切り分けを行った．また，同一の表現が複数の現象に分類される場合には，注目する現象毎に当該現象のみを取り除くような正規化を行う．例えば，「すまほ」という表現は，名詞の省略と異表記に分類されるべき表現であるが，これはそれぞれ「すまーとふぉん」「スマホ」に正規化されると期待される．なお，実際にはこのような複合はあまり見られなかった．

我々は，これらのタスクの結果を集約し，該当表現を正規化後の表現に置き換えることで，（原文, 正規化後文, アラインメント，参照訳）の 4 つ組としてデータセットを構築した．この際，同一の現象ラベルに属する表現が一文中に複数含まれる場合には，当該文が複数回評価されることによる過剰評価の影響を防ぐため，評価データから削除した．なお，「標準形」という概念が存在しない固有名詞に関しては，正規化タスクは行わず，原文とアラインメントを用いて評価を行うこととした．一連の手順により構築したデータセットの一例を表 5 に示す．また，各現象データセットの構成文数，ユニークな表現数，および正規化前後の文字単位による編集距離は表 6 のとおりであった．この結果から，口語表現や異表記は構成文数に対するユニークな表現の割合が高く，変化のパターンには多様性があることが窺える．また，口語表現は異表記に比べて正規化による編集距離が小さく，推論時の表層的な手がかりの利用可能性が品質の改善に寄与することが期待される．

4　翻訳モデル

本節では，構築したデータセットによる評価および分析の対象とした翻訳モデルについて説明する．我々は訓練データの規模や前処理の方法が異なる 5 種類のモデルを用意し，それぞれ 5 つの乱数シードで実験を行った．表 7 に各モデルの主な特徴，表 8 には学習に用いたデータセットの内訳を示す．また，実験は我々が独自に訓練したモデルに加え，Google 翻訳などの商用翻訳システムに対しても行った．実際にユーザからの入力を受け付けることで改善が行われ

1：	固有名詞
Orig. (Ja)	まじで**メルカリ**で野菜売ってる－！
Ref. (En)	There really are vegetables for sale at **Mercari**!
2：	名詞の省略
Orig. (Ja)	地味な**アプデ**だが
Norm. (Ja)	地味な**アップデート**だが
Ref. (En)	That's a plain **update** though
3：	口語表現
Orig. (Ja)	ここまで描いて飽きた、**かなちい**
Norm. (Ja)	ここまで描いて飽きた、**かなしい**
Ref. (En)	Drawing this much then getting bored, how **sad**.
4：	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 5　構築したデータセット中の文例．Orig., Norm., Ref. はそれぞれ，MTNT データセット中の原文，正規化タスク適用後の文，参照訳を表す．

データセット	構成文数	ユニークな表現数（割合）	編集距離（正規化前後）
固有名詞	943	747 (79.2%)	—
名詞の省略	348	234 (67.2%)	5.04
口語表現	172	153 (89.0%)	1.77
異表記	103	97 (94.2%)	3.42

表 6　各現象データセットの構成文数．

モデル	訓練データサイズ	分割手法	MeCab による発音化処理
Small	3.9 M	BPE （共有, 32k）	No
Large	14.0 M	BPE （共有 32k）	No
Char	14.0 M	文字 （共有）	No
Pron	14.0 M	文字 （共有）	Yes
Cat	28.0 M	BPE （共有, 32k）	Yes （一部）

表 7　各内製モデルの主な特徴．Large, Cat の両モデルには同一の BPE モデルを適用した．

データセット	サイズ	Small	Large, Char	Pron	Cat
TED talks	0.2 M	✓	✓		✓
KFTT	0.4 M	✓	✓		✓
JESC	3.2 M	✓	✓		✓
JParacrawl v2.0	10.1 M		✓		✓
（発音ベースコーパス）	14.0 M			✓	✓

表 8 訓練データセットの内訳.

る商用のシステムでは，UGC に対する頑健性がより高いことが期待される．そのようなシステムに対して実験を行うことで，より重要性の高い現象の存在を明らかにすると共に，質の高い翻訳を得るためにユーザが行うことのできる工夫としての正規化の有用性を確認する．

まず，訓練データの規模のみが異なる 2 つのモデルの比較を通して，ある現象が大規模なデータの収集により解決されうるかどうかを調査する．評価指標の差分による頑健性の評価に改善が見られない場合，前処理やモデリングにおいて特別な工夫を要する現象であり，より重要性が高いと考えることができる．Small モデルは WMT 2019 にて開催されたコンペティションにおいて訓練データとして公式に提供された TED talks, KFTT (**K**yoto **F**ree **T**ranslation **T**ask), JESC (**J**apanese-**E**nglish **S**ubtitle **C**orpus) の 3 つのデータセットを用いて学習したモデルである．一方で，Large モデルにはこれらのデータに加え，現在一般に利用可能な日英パラレルコーパスとして最大規模である JParacrawl v2.0 (Morishita et al. 2020) を用いた．また，データセットの前処理として，JESC にあたる部分の英語文がすべて小文字で記述されていたため，TED talks, KFTT, MTNT データセットを用いて学習した `moses` ツールキット (Koehn et al. 2007) の recaser を適用することで，文字種情報の復元を行った．さらに，文中の unicode 絵文字および顔文字を特殊トークンに置換し，後処理において復元を行った (Murakami et al. 2019)．また，翻訳時にコピーされることが望ましいユーザ名についても同様の処理を適用した[8]．トークンの単位にはサブワードを用い，`sentencepiece` (Kudo and Richardson 2018) により，両言語共有で語彙数が 32,000 となるように Byte-Pair-Encoding (BPE) を適用した (Sennrich et al. 2016)．

続いて，トークンの分割単位がモデルの頑健性に及ぼす影響を調べるため，Large モデルの訓練データに異なる分割を適用してモデルの学習を行った．Char モデルは，入出力を文字の系列と見なすことで，原言語側の一連の文字の並びから目的言語側の異なる文字の並びへの翻訳を行う (Wang et al. 2015)．数値やコードスイッチングを適切に扱うため，ここでも両言語の語彙は共有とした．先行研究において，文字ベースの翻訳モデルは BPE を用いたモデルに比べ

[8] @から始まる連続 15 文字以内の英数字の出現を正規表現により置換した．推論時には，MTNT データセットの情報源である Reddit の記法に倣い，*u/name* または *g/name* の形で表される表現をユーザ名と見なした．

誤植などの誤りに対する頑健性が高いことが示されている (Durrani et al. 2019)．我々は，UGC を取り扱う上でも分割手法の差異が翻訳品質に影響を及ぼすかについて改めて確認する．

さらに，我々のデータセットによる評価が，特定の現象に対する工夫を反映できることを検証するため，残る Pron，Cat の 2 モデルについては，原言語（日本語）側に特殊な前処理を施した．具体的には，日本語形態素解析ツールキット `MeCab` (Kudo et al. 2004) を用いて取得した各形態素の読みを結合することで発音ベースのコーパスを作成し，これを用いて学習を行った．日本語ではすべての単語の読みは表音文字であるひらがなまたはカタカナを用いて表すことができるが，Pron モデルの訓練には `MeCab` の標準出力であるカタカナに変換されたデータを用いた．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．中でも，異表記の定義に含まれるひらがな・カタカナの混同に適切に対処することを期待した．例えば，表 4 中の「アリガトウ」という表現は通常ひらがなで記述されるが，発音への変換を経ることで同一の表現（アリガトウ）に集約することが出来る．これにより，推論時に未知の表現が減少し，翻訳品質が改善すると考えられる．さらに，このような音声的情報は同音異義語の翻訳曖昧性解決にも有効であることが示されている (Liu et al. 2019)．

一方で，Cat モデルでは通常のコーパスと発音ベースのコーパスの結合コーパスを用いて学習を行った．コーパスの結合の方法には，それぞれのコーパス中の文を別々の文として縦方向に結合する方法と，`<sep>`などの特殊トークンを用いて同一文内で横方向に結合する方法が考えられるが，このモデルでは前者の方法を採用した[9]．また，発音ベースのコーパスに相当する部分をさらにひらがなに転写し，Large モデルと同一の BPE モデルを用いてサブワードに分割した．元文と発音に転写された文の 2 文から同一の目的言語文を出力するように学習を行う事で，一般的でない表記から生じる予期しない分割に対しても頑健な表現を得ることを期待した．

すべてのモデルには `fairseq` ツールキット (Ott et al. 2019) に実装された Transformer-base アーキテクチャを用い，ハイパーパラメータは Murakami et al. (2019) の設定に準じた．また，広く商用に利用されている機械翻訳システムとして，Google 翻訳[10]，および DeepL 翻訳[11]の 2 つのシステムの出力結果についても分析を行った[12]．

5　現象毎評価

本節では，構築した現象毎データセットを用いた評価によりニューラル機械翻訳システムの現状を概観する．具体的には，MTNT データセット中の原文と正規化後の文のそれぞれをモデ

[9] モデル名は Unix 系 OS における `cat` コマンドに由来する．特殊トークンによる `paste` モデルについても実験を行ったが，有意義な結果を得ることができなかったため本論文では省略する．
[10] `https://translate.google.co.jp`
[11] `https://www.deepl.com/translator`
[12] 本論文で用いた商用システムの翻訳結果は 2020 年 6 月 10 日現在のものである．

ルに入力し，単一参照訳 BLEU (Papineni et al. 2002; Post 2018)，およびアラインメントを用いた翻訳正解率の差分をもって評価を行う．それぞれの文は評価対象とする現象の有無の一点においてのみ異なるため，任意の評価指標によるスコアの差分の大きさは，当該現象がモデルに与える影響の大きさと考えることができる．この際，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いう直感のもと，差分を正規化後のスコアで除算した (Niu et al. 2020)．つまり，原文に対する翻訳を x_{orig}，正規化後の文に対する翻訳を x_{norm}，参照訳またはアラインメントを y，翻訳文と参照訳（アラインメント）を受け取りスコアを返す任意の評価指標を $\text{score}(x, y)$ としたとき，頑健性スコア ROBUST は以下のように定義される．

$$\text{ROBUST} = \frac{\text{score}(x_{\text{orig}}, y) - \text{score}(x_{\text{norm}}, y)}{\text{score}(x_{\text{norm}}, y)} \times 100 \tag{1}$$

頑健性スコア ROBUST は，正規化後の入力に対して到達可能な翻訳品質に対する相対的な品質の低下（上昇）と捉えることが出来る．我々は BLEU スコアのような文レベルの評価指標に加え，局所的な翻訳可能性を測定する正解率を相補的に用いることで，現状のモデルにおける改善点を明らかにするためのより詳細な分析を提供する．続く 5.1 節では定量評価，5.2 節では定性評価の結果を示す．

5.1　定量評価

内製モデル　表 9 および 10 にはそれぞれ，現象毎データセットに対する我々のモデルの BLEU スコア，当該箇所の翻訳正解率を示す．この結果から，訓練データ規模が小さい場合（SMALL モデル）は，正規形の存在する名詞の省略，口語表現，異表記のすべての現象において，原文

		SMALL	LARGE	CHAR	PRON	CAT
固有名詞	Orig.	10.88 ± 0.34	14.70 ± 0.26	12.60 ± 0.26	10.92 ± 0.20	14.08 ± 0.23
名詞の省略	Orig.	10.32 ± 0.39	14.14 ± 0.27	12.24 ± 0.41	10.50 ± 0.34	13.40 ± 0.40
	Norm.	10.76 ± 0.20	14.16 ± 0.30	12.16 ± 0.33	10.68 ± 0.31	13.62 ± 0.54
	ROBUST	−4.09	−0.14	+0.66	−1.69	−1.62
口語表現	Orig.	11.40 ± 0.36	13.60 ± 0.47	11.90 ± 0.42	10.34 ± 0.26	13.40 ± 0.45
	Norm.	12.12 ± 0.46	14.48 ± 0.19	12.22 ± 0.31	11.68 ± 0.92	14.34 ± 0.08
	ROBUST	−5.94	−6.08	**−2.62**	−11.47	−6.56
異表記	Orig.	9.96 ± 0.40	13.64 ± 0.32	13.40 ± 0.81	12.80 ± 1.26	14.80 ± 0.97
	Norm.	11.96 ± 1.02	15.70 ± 0.32	15.36 ± 0.55	13.32 ± 0.95	15.22 ± 1.49
	ROBUST	−16.72	−13.12	−12.76	**−3.90**	**−2.76**

表 9　我々のデータセットにおける BLEU スコアの測定結果（内製モデル）．* 5 シードの平均値 ± 標準偏差．Orig. は原文, Norm. は正規化後の文に対するスコアを表す．ROBUST は式 (1) による．

		Small	Large	Char	Pron	Cat
固有名詞	Orig.	34.80 ± 1.26	48.94 ± 0.32	47.40 ± 0.48	43.12 ± 0.71	47.96 ± 0.52
名詞の省略	Orig.	26.02 ± 1.17	35.66 ± 0.96	34.44 ± 1.30	31.10 ± 0.89	36.22 ± 1.33
	Norm.	30.76 ± 1.17	33.66 ± 1.29	34.02 ± 0.56	31.20 ± 0.87	32.66 ± 0.48
	Robust	−15.41	+5.94	+1.23	−0.32	+10.90
口語表現	Orig.	16.98 ± 0.76	13.84 ± 1.90	14.88 ± 1.57	10.90 ± 2.02	15.12 ± 1.75
	Norm.	22.90 ± 1.51	20.80 ± 0.55	20.24 ± 1.34	27.78 ± 1.43	31.88 ± 1.57
	Robust	−25.85	−33.46	**−26.48**	−60.76	−52.57
異表記	Orig.	13.68 ± 1.42	13.22 ± 1.86	12.44 ± 1.12	23.28 ± 1.82	28.34 ± 2.18
	Norm.	37.50 ± 2.64	38.26 ± 0.44	35.74 ± 1.29	31.66 ± 3.04	36.52 ± 1.46
	Robust	−63.52	−65.45	−65.19	**−26.47**	**−22.40**

表 10　我々のデータセットにおける正解率 (%) の測定結果（内製モデル）．* 5 シードの平均値 ± 標準偏差．Orig. は原文, Norm. は正規化後の文に対するスコアを表す．Robust は式 (1) による．

入力時に品質の低下が見られることがわかった．さらに，特に評価指標に正解率を用いた場合，Small モデルの名詞の省略に対する頑健性スコアが他のモデルに比べ著しく低いことが確認された．固有名詞や名詞の省略では，Large モデルにおいて原文入力時の正解率も大幅に向上したことから，訓練データを拡充しカバー率を向上することである程度対処可能な現象であると考えられる．しかし，一般に正規化により翻訳文の品質が低下することは考えにくく，正のスコアを示す頑健性スコアの解釈には疑問も残る．これに対して，我々は正規化後に翻訳文の品質が低下する例にはどのようなものがあるかを確かめるため，名詞の省略データセットの細分類を行った．その結果，省略形がアルファベットの頭字語である場合に，特に Large モデルで正規化前の入力に対する正解率が高く，それ以外の入力のパターンでは正解率の低下が確認された．これは，主に参照訳中にそのまま現れることの多い頭字語が，正規化により過剰に説明されてしまい表層の一致が取れなくなってしまったことに起因すると考えられる．しかし，実際には PC の翻訳が personal computer でも問題ないように，頭字語の翻訳には複数の正解が許容される場合がある．従って，正規化後の上界はより高いと考えられ Large モデルにおいても依然として課題は残されていると言える．

口語表現と異表記の 2 現象に注目すると，Large モデルは，BLEU スコアにおいて Small モデルよりも 2–3 ポイント程度優位であるものの，この改善は頑健性の向上に起因するものではないと考えられる．中でも，口語表現の正解率について，正規化後の入力にさえ Small モデルがより高い値を示したことは注目に値する．このことは，これらの現象が最先端の翻訳モデルにおいても十分に扱うことが難しく，データセットの拡充に加えて特別な対処を必要とすることを示唆している．

続いて，トークンの分割単位が各現象への頑健性に及ぼす影響を確認する．Large モデルと Char モデルに対する結果の比較から，文字ベースの分割は特に口語表現に対する頑健性を向上させることが確認できる．これは，3.2 節で述べたとおり，口語表現データセットの正規化前後の編集距離が他のデータセットに比べ小さいことに起因すると考えられる．先行研究における誤植と同様，周囲の文字が大きな手がかりとなるような現象に対しては，文字単位に分割することの有用性が確認できる．しかし，同時に BLEU スコアの絶対値や正規化後の表現に対する正解率において減少も見られることから，現象の存在が明らかである場合を除いては BPE に基づく分割が優位であるように思われる．一方で，異表記については一種の綴りの誤りと見なすことができるため，文字単位で扱うことによる頑健性の向上を期待したがその改善は限定的であった．これは，一般にある単語内の数文字の変化からなる誤植や口語表現に対して，単語単位で生じる異表記の問題としての難しさを示唆していると言える．

また，我々のデータセットによる評価が，特定の現象に対する頑健性に感受性を有するかについて，Pron モデルの結果から確認する．表 9 から，原文入力時の BLEU スコアに注目すると，一般に Pron モデルは流暢性において他のモデルよりも劣ることがわかる．一方で，このモデルにおいて頑健性の向上が期待される異表記のスコアに注目すると，同量のデータで学習した 3 モデル中では最小の −3.90，−26.47 ポイント（BLEU，正解率）となった．スコアが大きく改善した一因には，表音文字への統一により表現力が低下したことに起因する正規化後の正解率低下も考えられるものの，原文入力時の 10%に迫る改善は従来の自動評価では見いだせない解決策を切り拓く可能性の一例であると言える．

これに対して Cat モデルは，原文入力時の異表記の正解率について大幅な向上を達成しながら，同時に BLEU スコアの低下も抑えている点で Pron モデルの流暢性の問題を克服したモデルであると言える．正解率は全モデル中最大の 28.34%を示し，Large モデルからは 15%以上の飛躍が見られた（表 10）．また，口語表現に注目すると，BLEU スコアでは Large モデルから大幅な向上は見られないものの，正規化の適用により翻訳できる表現の割合に違いが生じていることがわかる．これは，3.2 節で述べたように口語表現と異表記における階層性に起因すると考えられ，同様に Cat モデルの異表記に対する頑健性を支持していると言える．このような改善が見られた理由としては，ひらがな化から生じる予期しない分割に対する適応可能性の向上があげられる．日本語では多くの高頻度の機能語は数文字のひらがなから構成される．そのため，ひらがな化された表現は単一の語としてではなく，より高頻度の機能語を生成するように分割されてしまう場合がある (Sasano et al. 2013)．発音に転写したコーパスの混合により，意図しない機能語を含む難しい系列からも正しい出力を行うように強いることで，共起関係がよりうまく捉えられたと考えられる．この結果から，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の考察を行う事が重要であることが窺える．

	BLEU				正解率 (%)			
	Google 翻訳		DeepL 翻訳		Google 翻訳		DeepL 翻訳	
	Orig. / Norm.	ROBUST	Orig. / Norm.	ROBUST	Orig. / Norm.	ROBUST	Orig. / Norm.	ROBUST
固有名詞	15.4 / —	—	16.0 / —	—	55.2 / —	—	50.5 / —	—
名詞の省略	14.6 / 15.0	−2.67	16.3 / 16.2	+0.62	41.1 / 36.8	+11.68	39.1 / 37.9	+3.17
口語表現	14.4 / 16.0	−10.00	15.6 / 15.8	−1.27	19.2 / 26.2	−26.72	22.7 / 28.5	−20.35
異表記	15.3 / 17.6	−13.07	14.4 / 15.2	−5.26	23.3 / 37.9	**−38.52**	18.4 / 35.0	**−47.43**

表 11　我々のデータセットにおける BLEU スコアおよび正解率の測定結果（商用システム）．* Orig. は原文, Norm. は正規化後の文に対するスコアを表す．ROBUST は式 (1) による．

商用システム　驚くべきことに，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，我々の異表記データセットに対する評価の結果からは改善の余地が見られることがわかった（表 11）．結果から，双方のシステムで原文入力時に比べ正解率で約 15%，BLEU スコアについても最大 2.3 ポイントの大きな低下が確認された．つまり，我々がユーザとしてこれらのシステムを用いる上では，複数の表記を許容できる表現についても，より一般的な表記に置き換える一工夫で，得られる翻訳の質を大きく高めることができると言える．また，BLEU スコアにおいて優れているシステムは必ずしも正解率で勝ってはいなかった．例えば，名詞の省略に対する DeepL 翻訳の正解率は Google 翻訳に比べ 2.0 ポイント低かったものの，同一のデータセットにおいて BLEU スコアでは 1.7 ポイント高い値を示した．これは，2 つのシステムが未知の表現に面した際の異なる挙動に由来すると考えられる．我々の実験では，DeepL 翻訳が一般的でない表現を省略することで全体の翻訳を流暢に保つのに対し，Google 翻訳はモデルの以後の出力に悪影響を及ぼすようなフレーズであっても何らかの出力を行う傾向が見られた．実用上は，あるシステムの適合率と再現率のどちらが重視されるべきかは応用先に依存する．現象毎の観点の提供に加え，適合率に基づく BLEU と再現率に基づく正解率の二面からの評価を行うことは，モデルの選択に大いに役立つと考える．

5.2　定性評価

本節では，モデルによって生成された出力が正規化の有無によりどのように変化したかを分析する．表 12 は我々のモデルによる実際の出力の一例である．

表中の例 (a) にはひらがなで記述された一般的でない表現「ぎゃくたい」が漢字（虐待）に置換された場合の出力の変化を示す．この例において，LARGE モデルでは原文を入力した際，誤って want to というフレーズを出力してしまった．このような出力となった理由として，当該表現が BPE モデルによって 4 つのトークンに過度に分割されていたことがあげられる．特に，しばしば願望を表す助動詞として用いられる「たい」という分割の存在によりこのような

(a) 異表記	
入力文	{ ぎゃくたい / 虐待 } だ
LARGE$_{orig}$	I want to do it!
CAT$_{orig}$	It's **abuse**!
LARGE$_{norm}$	It's **abuse**!
参照訳	It's abuse!
(b) 名詞の省略	
入力文	進化する { サバゲー / サバイバルゲーム }
LARGE$_{orig}$	The evolving mackerel game.
CAT$_{orig}$	Evolving sabage.
LARGE$_{norm}$	Evolving **Survival Game**
参照訳	An evolving survival game.
(c) 固有名詞	
入力文	平昌で「米日 VS 南北」の戦いが始まる
SMALL	In the Heisho era, the battle of 'South and South America' began.
LARGE	The Battle of 'America-Japan VS North-South' begins in **Pyeongchang**
参照訳	The "US and Japan vs. North and South Korea" battle has begun in Pyeongchang.

表 12 内製モデルによる実際の出力例．*{ 正規化前 / 正規化後 }

誤訳が生じたと推測される．一方で，CAT モデルでは入力文は同一の BPE モデルで分割されているにも関わらず，ひらがなの入力に対しても正しい訳である abuse を出力することができた．多くの場合，助動詞「たい」に先行する文字は「したい」や「食べたい」のように「い」または「え」の子音で構成される．発音ベースのコーパスを加えることにより，願望の意味で用いられない熟語中の「たい」などの負例が十分数出現したため，出力が改善したと考えられる．異表記は日本語のように多様な文字を用いる言語に特有な現象であるものの，類似の問題はアルファベットを用いる言語にも存在する．例えば (a) の例は，単語の一部が既存の他の単語に連想されることで生じた誤りであるが，誤植の一部では同様の問題が生じる．そのような場合に，その表現が誤りであることを文脈から正しく判断し，正確な翻訳を行う事は現在の翻訳システムにとって大きな課題である．

次の例 (b) は名詞の省略データセット中の例である．この例では，LARGE モデルは原文中の表現「サバゲー」を，誤って「サバ」と「ゲー」の二語の組みあわせとして翻訳している．一方で CAT モデルでは，同様に正しく翻訳することはできなかったものの，対象の表現をアルファベットへ転写することで対処した．この結果から，CAT モデルはサバとゲームが共起しにくいといった情報をよりうまく捉えているように思われる．また，データセット中には「生保」の

ようにある表現が文脈に応じて複数の表現の省略となりうる例[13]も存在した．こういった曖昧性に対してモデルをさらに頑健にするためには，単語や文内に限らずより広い文脈を考慮することが求められていると言える．

最後に (c) の例では，文中の「平昌」という固有名詞を Small モデルは扱うことができなかったのに対し，Large モデルは正しく翻訳することができた．これは，訓練データを新たに加えたことでテストデータ中に既知の固有名詞が増加したことに由来すると考えられるが，我々は訓練データの規模の増大が固有名詞に対するモデルの頑健性を説明する理由として必ずしも十分ではないと考えている．その理由として，固有名詞にはある時期を境とした出現頻度の変化が生じやすいことがあげられる．例えば，「平昌」という表現を例にあげると，これは同所でオリンピックが開催された 2018 年前後の記事を含むコーパスでは出現数が大幅に増加することが見込まれる．Small モデルの訓練に使用されたコーパスは 2018 年以前に作成されたものであることから，当該表現の相対的な出現数の少なさが翻訳を困難にしたと考えられる．我々は，日々新たに生まれる未知の固有名詞に対して真に頑健なシステムを評価するためには，評価データが訓練データよりも常に時系列的に新しくなるように分割を行うべきだと考える．しかし，現実的には入手可能ないかなる訓練データよりも新たなデータを作成することは不可能である．そのため，このような要因の存在を認識し留意した上で，我々のデータセットを用いた評価を固有名詞に対する一定の指針として用いることは有意だと考える．

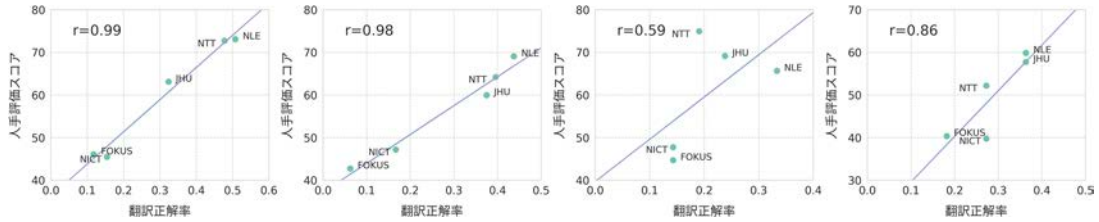
6　考察

6.1　WMT コンペティション参加システムの現象毎分析

現象毎データセットの潜在的な活用事例を探るため，我々は WMT 2019 にて行われた機械翻訳システムの頑健性に関するコンペティションに投稿されたシステム出力の現象毎分析を行った．コンペティションの公式サイト[14]より，MTNT データセットのブラインドセットにあたる部分について 5 つのシステムの出力をダウンロードした．その後，構築したデータセットよりブラインドセットに由来する部分を抽出し，評価の対象となる固有名詞 136 文，名詞の省略 48 文，口語表現 21 文と異表記 11 文を得た．システム出力には，それぞれの文に対し 3 名の翻訳者による人手評価で 1（非常に悪い）– 100（非常に良い）のスコアが付与されている (Li et al. 2019)．我々は，3 名のスコアの平均を翻訳文のスコアとし，各現象データセットに対して平均を取ることで，モデルのある現象に対する人手評価スコアを算出した．

図 6 には，各現象に対する人手評価スコアと翻訳正解率の相関を示す．図より，特に固有名

[13] 生命保険，または生活保護．
[14] http://matrix.statmt.org/matrix/systems_list/1917



(a) 固有名詞　(b) 名詞の省略　(c) 口語表現　(d) 異表記

図 6　WMT コンペティション投稿システム出力に対する人手評価スコアと翻訳正解率の相関．r はピアソンの積率相関係数を表す．

詞と名詞の省略について人手評価と正解率の間にはピアソンの積率相関係数 $r > 0.9$ の非常に強い正の相関があることがわかった．正解率という，文中の局所的な翻訳可能性に注目した指標が文全体の翻訳品質と高い相関を示したことは注目に値する．これは，人間が翻訳の良さを判断する上で特定の表現をより重点的に考慮する可能性を示唆している．ひとつの理由としては，人間にとって翻訳文中に固有名詞が含まれるかどうかの判断は容易であることがあげられる．こうした表現の訳抜けは，他の表現の訳抜けに比べて評価への影響が大きいと考えられる．一方で，今回比較したシステムは個々の性能差も大きいことから，相関が現象に関連する部分のみに起因すると帰結することは難しい．今後の方向性としては，類似の性能を示すモデルを用いて同様の検証を行うとともに，他の評価指標との相補的な利用を模索し，よりよい人手評価の近似について検討することが考えられる．

7　おわりに

本研究では，日英機械翻訳システムの精緻な評価に向けた第一歩として，ユーザ生成コンテンツに着目した言語現象毎評価データセットを提案した．具体的には，固有名詞，名詞の省略，口語表現，異表記の 4 つの現象に対するモデルの頑健性について評価・分析を行った．

分析の結果，広く商用に利用される翻訳システムを含む多くのモデルが，異表記を含む入力に対して不安定であることがわかった．これは，現在の主流であるモデルや訓練データの大規模化が異表記のような特異な入力に対する頑健性の十分条件ではなく，異文化・多言語交流を促進する翻訳システムの構築に向けては特別な工夫を要することを示唆している．

また，過去に開催されたコンペティションにおけるシステムの人手評価スコアを用いて，現象毎に人手評価と翻訳正解率の相関関係を調査した．実験の結果，特に固有名詞と名詞の省略について両者の間に非常に強い正の相関が見られたことから，従来の適合率に基づく評価に加えて，我々の評価を用いることで人手評価をよりよく近似する可能性を確認した．

我々は，機械翻訳システムのさらなる発展のため構築したデータセットを公開している[15]．我々のデータセットが，より実用的な機械翻訳システムの構築に向けて，機械翻訳コミュニティを一歩前進させるための先駆けとなることを期待する．

謝　辞

本論文の内容の一部は，The 28th International Conference on Computational Linguistics (COLING 2020) で発表されたものです (Fujii et al. 2020)．また，本研究の一部は JSPS 科研費 JP19H04162，JP20J21694 の支援を受けて行いました．本論文の執筆にあたり，有益なコメントを頂きました査読者，担当編集委員の皆様に感謝申し上げます．

参考文献

Anastasopoulos, A., Lui, A., Nguyen, T. Q., and Chiang, D. (2019). "Neural Machine Translation of Text from Non-Native Speakers." 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. 3070–3080.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (2019). "Findings of the 2019 Conference on Machine Translation (WMT19)." In *Proceedings of the 4th Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 1–61.

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018). "Evaluating Discourse Phenomena in Neural Machine Translation." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1304–1313.

Belinkov, Y. and Bisk, Y. (2018). "Synthetic and Natural Noise Both Break Neural Machine Translation." In *6th International Conference on Learning Representations, ICLR 2018*.

Berard, A., Calapodescu, I., Dymetman, M., Roux, C., Meunier, J.-L., and Nikoulina, V. (2019a). "Machine Translation of Restaurant Reviews: New Corpus for Domain Adaptation and Robustness." In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pp. 168–176.

[15] `https://github.com/cl-tohoku/PheMT`

Berard, A., Calapodescu, I., and Roux, C. (2019b). "Naver Labs Europe's Systems for the WMT19 Machine Translation Robustness Task." In *Proceedings of the 4th Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 526–532.

Durrani, N., Dalvi, F., Sajjad, H., Belinkov, Y., and Nakov, P. (2019). "One Size Does Not Fit All: Comparing NMT Representations of Different Granularities." 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. 1504–1516.

Freitag, M., Grangier, D., and Caswell, I. (2020). "BLEU Might Be Guilty but References Are Not Innocent." *arXiv*, **abs/2004.06063**.

Fujii, R., Mita, M., Abe, K., Hanawa, K., Morishita, M., Suzuki, J., and Inui, K. (2020). "PheMT: A Phenomenon-wise Dataset for Machine Translation Robustness on User-Generated Contents." In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5929–5943.

Graham, Y., Haddow, B., and Koehn, P. (2019). "Translationese in Machine Translation Evaluation." *arXiv*, **abs/1906.09833**.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., and Zhou, M. (2018). "Achieving Human Parity on Automatic Chinese to English News Translation." *arXiv*, **abs/1803.05567**.

Heigold, G., Varanasi, S., Neumann, G., and van Genabith, J. (2018). "How Robust Are Character-Based Word Embeddings in Tagging and MT Against Wrod Scramlbing or Randdm Nouse?" In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pp. 68–80.

Ikeda, T., Shindo, H., and Matsumoto, Y. (2016). "Japanese Text Normalization with Encoder-Decoder Model." In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pp. 129–137.

Isabelle, P., Cherry, C., and Foster, G. (2017). "A Challenge Set Approach to Evaluating Machine Translation." In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2486–2496.

Junczys-Dowmunt, M. (2018). "Dual Conditional Cross-Entropy Filtering of Noisy Parallel Corpora." In *Proceedings of the 3rd Conference on Machine Translation: Shared Task Papers*, pp. 888–895.

Karpukhin, V., Levy, O., Eisenstein, J., and Ghazvininejad, M. (2019). "Training on Synthetic Noise Improves Robustness to Natural Noise in Machine Translation." In *Proceedings of the*

5th Workshop on Noisy User-generated Text (W-NUT 2019), pp. 42–47.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). "Moses: Open Source Toolkit for Statistical Machine Translation." In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 177–180.

Koehn, P., Khayrallah, H., Heafield, K., and Forcada, M. L. (2018). "Findings of the WMT 2018 Shared Task on Parallel Corpus Filtering." In *Proceedings of the 3rd Conference on Machine Translation: Shared Task Papers*, pp. 726–739.

Krippendorff, K. (2011). "Computing Krippendorff's Alpha-Reliability.".

Kudo, T. and Richardson, J. (2018). "SentencePiece: A Simple and Language Independent Subword Tokenizer And Detokenizer for Neural Text Processing." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71.

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). "Applying Conditional Random Fields to Japanese Morphological Analysis." In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 230–237.

Landis, J. R. and Koch, G. G. (1977). "The Measurement of Observer Agreement for Categorical Data." *Biometrics*, **33** (1), pp. 159–174.

Li, X., Michel, P., Anastasopoulos, A., Belinkov, Y., Durrani, N., Firat, O., Koehn, P., Neubig, G., Pino, J., and Sajjad, H. (2019). "Findings of the First Shared Task on Machine Translation Robustness." In *Proceedings of the 4th Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 91–102.

Liu, H., Ma, M., Huang, L., Xiong, H., and He, Z. (2019). "Robust Neural Machine Translation with Joint Textual and Phonetic Embedding." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3044–3049.

Luong, T., Pham, H., and Manning, C. D. (2015). "Effective Approaches to Attention-based Neural Machine Translation." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421.

Michel, P. and Neubig, G. (2018). "MTNT: A Testbed for Machine Translation of Noisy Text." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 543–553.

Morishita, M., Suzuki, J., and Nagata, M. (2020). "JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus." In *Proceedings of The 12th Language Resources and*

Evaluation Conference, pp. 3603–3609.

Murakami, S., Morishita, M., Hirao, T., and Nagata, M. (2019). "NTT's Machine Translation Systems for WMT19 Robustness Task." In *Proceedings of the 4th Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 544–551.

Nagase, T., Tsukada, H., Kotani, K., Hatanaka, N., and Sakamoto, Y. (2011). "Automatic Error Analysis Based on Grammatical Questions." In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, pp. 206–215.

Napoles, C., Sakaguchi, K., and Tetreault, J. (2017). "JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction." In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 229–234.

Niu, X., Mathur, P., Dinu, G., and Al-Onaizan, Y. (2020). "Evaluating Robustness to Input Perturbations for Neural Machine Translation." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8538–8544.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). "fairseq: A Fast, Extensible Toolkit for Sequence Modeling." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 48–53.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). "BLEU: a Method for Automatic Evaluation of Machine Translation." In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318.

Post, M. (2018). "A Call for Clarity in Reporting BLEU Scores." In *Proceedings of the 3rd Conference on Machine Translation: Research Papers*, pp. 186–191.

Rosca, M. and Breuel, T. (2016). "Sequence-to-sequence Neural Network Models for Transliteration." *arXiv*, **abs/1610.09565**.

斉藤いつみ, 貞光九月, 浅野久子, 松尾義博 (2014). 正規-崩れ文字列アライメントと文字種変換を用いた崩れ表記正規化に基づく日本語形態素解析. 言語処理学会第 20 回年次大会 発表論文集, pp. 777–780. [I. Saito et al. (2014). Seiki-Kuzure Mojiretsu Araimento to Mojishu Henkan wo Mochiita Kuzurehyoki Seikika ni Motozuku Nihongo Keitaiso Kaiseki. Proceedings of the 20th Annual Meeting of the Association for Natural Language Processing].

Sasano, R., Kurohashi, S., and Okumura, M. (2013). "A Simple Approach to Unknown Word Processing in Japanese Morphological Analysis." In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pp. 162–170.

Sennrich, R. (2017). "How Grammatical is Character-level Neural Machine Translation? Assess-

ing MT Quality with Contrastive Translation Pairs." In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 376–382.

Sennrich, R., Haddow, B., and Birch, A. (2016). "Neural Machine Translation of Rare Words with Subword Units." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725.

Shao, Y. and Nivre, J. (2016). "Applying Neural Networks to English-Chinese Named Entity Transliteration." In *Proceedings of the 6th Named Entity Workshop*, pp. 73–77.

Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). "Evaluating Gender Bias in Machine Translation." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1679–1684.

Ugawa, A., Tamura, A., Ninomiya, T., Takamura, H., and Okumura, M. (2018). "Neural Machine Translation Incorporating Named Entity." In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3240–3250.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). "Attention is All you Need." In *Advances in Neural Information Processing Systems 30*, pp. 5998–6008.

Wang, L., Isabel, T., Chris, D., and Alan W., B. (2015). "Character-based Neural Machine Translation." *arXiv*, **abs/1511.04586**.

White, J. S., O'Connell, T. A., and O'Mara, F. E. (1994). "The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches." In *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas*, pp. 193–205.

略歴

藤井　　諒：2019 年東北大学工学部電気情報物理工学科を卒業，同年より東北大学情報科学研究科博士前期課程に在籍中（2021 年 3 月修了予定）．多言語処理に関心があり主に機械翻訳に関する研究に従事している．

三田　雅人：理化学研究所革新知能統合研究センター自然言語理解チームテクニカルスタッフ．2016 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．日本マイクロソフト株式会社のエンジニアを経て，2018 年より現職．同年，東北大学情報科学研究科博士後期課程に進学．文法誤り訂正を中心とした自然言語処理による言語学習／教育支援に関心がある．言語処理学会，ACL 各会員．

阿部香央莉：2020 年東北大学大学院情報科学研究科博士前期課程修了．現在，

東北大学情報科学研究科博士課程取得に向け研究を進めている．2020 年度日本学術振興会 DC1 採択．多言語処理を伴う機械翻訳や言語類型学的観点からの言語処理分析に関心がある．言語処理学会，ACL 各会員．

塙　　一晃：理化学研究所革新知能統合研究センター自然言語理解チームテクニカルスタッフ．2019 年東北大学大学院情報科学研究科博士前期課程修了．同年より現職．2020 年東北大学大学院情報科学研究科博士後期課程に進学．主に自然言語処理に関する研究に従事．言語処理学会，ACL 各会員．

森下　　睦：NTT コミュニケーション科学基礎研究所研究員．2017 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．同年より現職．主に機械翻訳，言語資源構築に関する研究に従事．言語処理学会，情報処理学会，ACL 各会員．

鈴木　　潤：2001 年から 2018 年まで日本電信電話株式会社コミュニケーション科学基礎研究所研究員（主任研究員／特別研究員）．2005 年奈良先端科学技術大学院情報科学研究科博士後期課程修了．現在，東北大学データ駆動科学・AI 教育研究センター教授．

乾　健太郎：東北大学大学院情報科学研究科教授．1995 年東京工業大学大学院情報理工学研究科博士課程修了．同大学助手，九州工業大学助教授，奈良先端科学技術大学院大学助教授を経て，2010 年より現職．2016 年より理化学研究所 AIP センター自然言語理解チームリーダー兼任．情報処理学会論文誌編集委員長，同会自然言語処理研究会主査，言語処理学会論文誌編集委員長等を歴任，2020 年より言語処理学会副会長．

（2020 年 10 月 27 日　受付）
（2020 年 12 月 28 日　再受付）
（2021 年 2 月 4 日　採録）