機械翻訳に対する文間文脈を考慮した評価と分析

長我部 恭行　甲斐 優人　石井 奏人
荻野 天翔　黒澤 道希　小町 守
首都大学東京
{osakabe-takayuki, kai-hiroto, ishii-kanato, ogino-ozora,
kurosawa-michiki}@ed.tmu.ac.jp, komachi@tmu.ac.jp

1 はじめに

ニューラル機械翻訳の研究が盛んになり，統計的機械翻訳と比較し流暢な出力が得られるようになった [5]．しかし，現在行われている研究の多くは単文単位での翻訳精度の改善を目指しているものであり，前後の**文間文脈**は考慮されていない．[*1]加えて，現在の機械翻訳の評価においては単文単位もしくはシステム単位での評価が一般的である．

一方，翻訳における品質は単文単位では測れない部分も多く存在する．例えば指示語が前の文の内容を指す場合や同じ語（特に専門用語）が同じ表現で記述されているかなどである．近年ではこれらを考慮すべく，文間文脈を考慮した新たなモデルも提案されつつある．[1][3] 本論文ではこのような翻訳を**文脈翻訳**と呼ぶ．

そこで我々は翻訳に対して文脈を考慮した評価を人手で行った．今回は文章を対象とすると考慮すべき文数が増大してしまうため，段落に限定して評価する．特に，単文単位の評価と段落内の文脈を考慮した評価における流暢性・妥当性の違いと段落単位での一貫性，および段落単位での流暢性・妥当性について評価した．また，それらの評価を分析することにより文脈を考慮した翻訳を行う際に評価を下げうるポイントを考えた．また，公開したデータを活用することにより，翻訳の正しさについて評価する一つの目安とすることができる．

本論文の貢献は以下の通りである．

- 単文単位での翻訳結果を用いて，単文および段落単位での評価を人手で行い，データを公開した[*2]．
- 評価結果を用いて，文脈を考慮した翻訳を行う際に優先すべきポイントを分析した．

2 関連研究

2.1 文脈翻訳

前後文を考慮したニューラル翻訳は近年複数行われている．それら複数のモデルに対して，Muller ら [3] は実験設定を統一して分析を行っている．彼らは，実験設定によって BLEU スコアなどに差はあるものの，文脈を考慮した翻訳の方が品質の高い出力を得られることは一貫していると述べている．

2.2 機械翻訳評価

機械翻訳の評価に一般的に用いられている指標は Papineni らが発表した BLEU [4] である．この指標は文の n-gram 一致率により評価する指標であり，利便性が優れているためにほとんどの論文で用いられている．しかしながら，表層しか用いないこの指標では適切な評価ができない．[2] 特に前後文を考慮する必要のある文脈翻訳に対しては一貫性などの重要な素性が考慮されていないため，不当に高くもしくは低くスコアを出す可能性も考えられる．

その一方で，文脈翻訳を適切に評価する指標は確立されていない．現在の文脈翻訳の論文では BLEU などの単文に対する指標を用いるか人手による評価が行われている．

3 翻訳評価アノテーション

3.1 評価に用いたデータ

本研究では翻訳のために文間文脈の情報が必要と考えられる報道文章に対してアノテーションを行うことにし，OPUS で公開されている “Global Voices” を利用した[*3]．

公開されている Global Voices の En-Ja のデータからランダムに 63 段落，222 文を抽出しアノテーションを行った．なお本論文では，元データに付けられた自動文アライメントの結果を元に，英日ともに 2 文以上含まれている段落に限定して選択するものとした．

OPUS のデータは人手による日本語訳がすでに付与されているが，今回はシステムに対する評価を行うために Google 社による NMT システム[*4]と SMT システ

[*1] 本論文では前後文を含む文間文脈を文脈と定義する．
[*2] https://github.com/tmu-nlp/MTEval4GV
[*3] http://opus.nlpl.eu/GlobalVoices.php
[*4] https://translate.google.co.jp/?hl=ja

表 1　流暢性の評価基準および例文

評価	基準	例文
5	日本語として完璧である	締め切りは 2007 年 12 月 21 日です。
4	日本語母語話者が言っていても違和感がない程度	情報は多くの情報源から来ている。
3	非日本語母語話者が言っていても違和感がない程度	オーランド・カストロは、国際社会はアフリカを気にしないと主張している。
2	日本語の構文として不適切である部分が多い	観光は、」フェズからの眺めを書き込みます。
1	理解不能である	彼らはさようなら、“トルコのボーイフレンド、そう”

ム[*5]の 2 システムによる翻訳を新たに取得し[*6]，独自に評価した．そのため，今回の評価は 63 段落，222 文に対し 2 システム分のデータ量が存在する．

3.2　アノテーション環境

前述のデータに対して，情報系の大学生 4 名でアノテーションを行った．最初に単文単位でのアノテーションを行い，その後に段落単位でのアノテーションを行った．なお，それぞれのアノテーションでは評価に必要な最低限の情報のみ[*7]を表示し，前後文（段落）を見られないようにランダムな順番で評価を行った．

また，どちらも全体のアノテーションを行う前に，評価の基準を統一するために，単文単位では 2 システム合わせて 100 文程度，段落単位では 2 システム合わせて 20 段落程度を用いて基準に関するすり合わせ[*8]を行い，基準を決めたのちに残りのデータに対してそれぞれが評価を行った．なお，設定した基準については次節で述べる．

3.3　アノテーション基準

3.3.1　単文単位でのアノテーション

単文単位でのアノテーションは，単文ごとの流暢性，妥当性をそれぞれ最低点の 1 から最高点の 5 の 5 段階で評価した．

流暢性の評価基準と例文を表 1 に示す．

次に，妥当性の評価基準と例文を表 2 に示す．また，評価する際，原文中の重要度を考慮して評価し，不一致の割合のみで評価しない．

3.3.2　段落単位でのアノテーション

段落単位でのアノテーションは，まず段落全体を見ながら単文単位の妥当性を見直した．また新たに段落を通して文末表現や翻訳の統一感を測る一貫性，段落全体での流暢性，妥当性の 3 つを評価した．

単文単位の妥当性は，段落ごとに文の前後関係を見て新しく評価を付け直した．基準は表 2 に則っている．

段落単位の流暢性，妥当性は，表 1，表 2 の基準に従って段落全体に対して最低点の 1 から最高点の 5 の 5 段階で評価した．

一貫性は最高点を 5 点とし，以下に示す条件に当てはまるたびに減点する方式で評価した．

- 文末表現が統一されていない場合

  「です・ます」調と「だ・である」調が混在している

- 文章の大意に影響を与えるような単語の表記の揺れがある場合

  “state language” を「状態言語」と「国家言語」に翻訳している

4　分析

4.1　単文単位の流暢性，妥当性の定量的分析

アノテーションの一致度を測る統計量として今回は Fleiss の Kappa と Kendall の一致係数を算出した．

Fleiss の Kappa は 2 人以上の複数の評価者間でアノテーションが完全に一致している度合いを測る統計量で 0 から 1 の値を取り，値が大きいほど一致度が高い．Kendall の一致係数は評価者間のアノテーションの関連性を表す統計量で 0 から 1 の値を取り，こちらも値が大きいほど関連性が高い．

単文単位でのアノテーションの流暢性，妥当性の両者について Fleiss の Kappa と Kendall の一致係数を表 3 に示す．Fleiss の Kappa については流暢性，妥当性ともに 0.25 付近であり，一致率が高いとは言えない．しかし Kendall の一致係数では 0.6 を超えた高い数値を示しており，評価者間に大幅な乖離はなく，評価はかねがね一致しているといえる．

単文のみを見て評価した妥当性と段落全体を見て評価した妥当性の評価数を表 4 に示す．評価者 4 人中全員が 5 と評価した数が減少し，また半分の評価者が 4 と評価した数が減少して 3 以下の評価の数が増加した．つまり，段落全体を見ることで単文単位では気づかなかった間違いを指摘できるようになったことから，文間文脈が妥当性判断の重要な要因になると考えられる．

[*5] Google スプレッドシートの translate 関数を用いた．
[*6] 2018 年 10 月 25 日取得
[*7] 流暢性を判断する際には日本語文のみ，妥当性を判断する際には英語文と日本語文の両方
[*8] 評価が 2 以上離れたものを対象として基準をすり合わせた．

表 2　妥当性の評価基準および例文

評価	基準	原文	翻訳文
5	原文を間違いなく翻訳している	Some bloggers also publish their thoughts, short stories, and poems in their blogs.	ブロガーの中には、彼らの思考、短編小説、および詩をブログに掲載するものもあります。
4	原文の大意を間違いなく翻訳している	It all ended well when the cops decided to compare our ID numbers.	警察が私たちの ID 番号を比較することを決めたとき、それはすべて終わった。
3	原文の大意は間違えていないが，それ以外の部分を間違って翻訳している	While Super Tuesday has come and gone in the U.S., conversations carry on in its wake among bloggers in the booming Japanese blogosphere.	スーパー火曜日が米国に来てしまったが、会話が活況を呈し日本のブロゴスフィアでブロガーの間でそのきっかけに続けていきます。
2	原文の大意は間違っているが，部分的に翻訳できている	The Tokyo Central Post Office building was designed by Yoshida Tetsuro, a Japanese modern architect who also designed other buildings commissioned by the Ministry of Communications, and was completed in 1931.	東京中央郵便局の建物は吉田鉄郎、また通信省の委託を受け、他の建物を設計し、1931 年に完成した日本の近代建築家によって設計されました。
1	原文と翻訳文の意味が全く異なる	存在せず	存在せず

表 3　単文単位の流暢性と妥当性の Fleiss の Kappa と Kendall の一致係数

	流暢性	妥当性
Kappa 係数	0.277	0.246
Kendall の一致係数	0.768	0.675

表 4　単文単位での妥当性と段落単位を見てから評価した妥当性の評価数

	評価者 1	評価者 2	評価者 3	評価者 4
1	14 → 14	4 → 3	0 → 0	3 → 7
2	47 → 54	47 → 55	52 → 57	49 → 54
3	150 → 163	153 → 162	167 → 166	116 → 124
4	158 → 150	137 → 139	173 → 173	149 → 146
5	75 → 63	103 → 85	52 → 48	127 → 113

4.2　単文単位での分析

流暢性．　単文の流暢性において，評価者の評価をもとに分析し，評価が一致しない要因を 2 つに分類した．まず，前後文との関係を持つ単文を評価するとき，評価が一致しなかった．例えば，“スーダンも.” という単文を評価するとき，文章が途切れており未完成の文に見える．しかし，前文が存在している場合，この文は前文の部分を省略していると捉えられる．次に，単語単位もしくは文節単位で評価した場合，流暢性は高いが，繋げて読むと単文自体の流暢性は低い場合が多い．例えば，「ダルフール紛争は、少なくとも 20 万人の命を主張し 200 万難民や避難民が発生しています」の単文の文節は，「200 万難民」以外の文節の流暢性は高い．しかしながら，繋げて読むと違和感があるため，文自体の評価は 3 以下になる．このように，1 つの文節が評価に対してどの程度影響を与えるかによって単文全体の評価が揺れる．

妥当性．　主に単語の捉え方により妥当性の評価が異なることがわかった．まず，単語の評価者間の解釈の違いにより評価が揺れた．例えば，“By” の訳し方において，原文が属する分野によって違ってくる．“By” は書籍では著者，インターネット上では投稿者を表すと考えられるが，単文からはどの分野に属するかを判断することができないことにより，評価者間で評価が異なった．次に，単語が固有名詞として判別されるか否かにより，評価の揺れがあった．例えば，SNS から引用される場合にはユーザ名が併記される場合がほとんどである．これらは固有名詞（人名）として扱われるべきものであるが，「韓国の大統領府、青瓦台（@BlueHouseKorea は）国民感情を落ち着かしようとしています。」のような文において「青瓦台」をアカウント名として固有名詞であると捉えるかによって評価が異なった．

4.3　段落単位での分析

段落単位では主に流暢性に関するもの，妥当性に関するもの，一貫性に関するもの，またはそれ以外と分類して分析した．

流暢性．　最初に流暢性に関わるものとして，段落の中に一文でも流暢性の低い文が混ざっていると段落全体での流暢性が低くなる．段落内の文数が多い場合，日本語母語話者の発言として違和感がある文が一文でも存在すると，割合的には小さいが違和感を感じるためだと思われる．

妥当性．　次に段落単位の妥当性については文数が多い段落は妥当性の低い文が混ざっていたとしてもそれに影響されて低く評価されることはなかった．これは段落全体を見ることで単文単位では妥当性の低い文についても前後の文から内容を推測し，補うことができるからだと思われる．

表 5　ある文の翻訳の仕方により評価が揺れる例

文数	原文	翻訳文 (SMT)	翻訳文 (NMT)
1	The conflict in Darfur has claimed at least 200 thousand lives and caused two million refugees and displaced persons.	ダルフール紛争は、少なくとも 20 万人の命を主張し 200 万難民や避難民が発生しています。	ダルフールの紛争は少なくとも 20 万人の命を奪い、2 百万人の難民と避難民をもたらした。
2	What about Rwanda, Ethiopia, Somalia?	ルワンダ、エチオピア、ソマリアはどうですか？	何ルワンダ、エチオピア、ソマリアは？
3	But does that matter?	しかし、それは問題ではありませんか？	しかし、それは重要ですか？
4	Important, indeed, is the wedding today, at the Elys_e Palace, of the French president, Nicolas Sarkozy, and singer and former model, Carla Bruni.”	重要なのは、確かに、結婚式はフランス大統領、サルコジ、歌手、元モデル、カーラ・ブルーニの、エリゼ宮で、今日です。」	重要なのは、フランスの大統領、ニコラス・サルコジ、そして歌手で元モデルのカーラ・ブルニの、エリゼ宮殿での結婚式です。

一貫性．　さらに，一貫性に関するものとして，ある文の翻訳の仕方によりその後の文のとらえ方が変化し，評価が揺れる例が存在した．その例文を表 5 に示す．この例において，3 文目の翻訳の仕方によりその後の 4 文目が段落全体として話が一貫しているのかどうかが変化する．SMT の翻訳文では前半のダルフール紛争の話題から突然の問いかけに違和感があり，さらに 4 文目の結婚式の話も話題の一貫性が感じられない．しかし，NMT の翻訳文では前半のダルフール紛争の話題に対してそれが重要かどうかを問いかけ，8 文目の返答は評価者によっては自然に感じることができ，SMT の翻訳文よりも NMT の翻訳文のほうが一貫性が保てていると評価される場合もある．

4.4　段落を考慮した単文評価の妥当性

単文単位で評価した妥当性と段落を見たうえで新たに評価しなおした妥当性について，評価が変化するものを 2 つに分類した．

文間文脈．　例えば “What is interesting to note is that the Chinese also depend on themselves to sell their goods away from the stores.” という文の翻訳が「興味深いのは、中国人も店舗から商品を売ることに依存しているということです。」と翻訳される場合，単文で見たときには “away from” の翻訳が間違っていると感じるだけである．しかし，この後ろの翻訳文が「今はいつも、中国のセールスマンが訪問しています。」となっているためこの文は「店舗 “から離れて” 商品を売る」と翻訳されるべきであるとわかるため，元の妥当性の評価より評価を下げるべきであることがわかる．

多義語．　単文の場合，多義語を執筆者の意図した意味で翻訳されていなくても違和感を感じることは少ないが，段落全体を見た後に翻訳の間違いが明らかになるということである．例えば，“It’s easy to apply.” という原文に対して「それは簡単に適用できます。」という翻訳がついていた場合，単文では全く違和感のない翻訳になっていて妥当性は高い評価になるが，後の文で “Anyone can apply.” を「誰でも申し込むことができます。」と翻訳していることから，“apply” を「適用する」ではなく「申し込む」と翻訳することが正しいと判断することができる．

5　おわりに

本研究では，OPUS で公開されている “Global Voices” を利用して機械翻訳に対する文脈を考慮した評価，および分析を行った．本研究で設けたアノテーション基準では，原文が原因の翻訳の間違いを評価することができなかったので，原文を考慮した上で妥当性や一貫性を評価する評価指標を設けることで，原文が原因の翻訳についても分析が可能であるかについて検討していきたい．さらに，本研究の分析を元に段落を超えた文書単位で文脈を考慮したアノテーション，分析を検討したい．

参考文献

[1] Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. Evaluating discourse phenomena in neural machine translation. In *Proc. of NAACL*, pp. 1304–1313, 2018.

[2] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluation the role of BLEU in machine translation research. In *Proc. of EACL*, pp. 246–256, 2006.

[3] Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proc. of WMT*, pp. 61–72, 2018.

[4] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*, pp. 311–318, 2002.

[5] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. Modeling coverage for neural machine translation. In *Proc. of ACL*, pp. 76–85, 2016.