ニュースのライティングスタイルの差分を考慮した日英機械翻訳のテストセットの開発

衣川和尭　美野秀弥　後藤功雄　山田一郎
NHK 放送技術研究所
{kinugawa.k-jg,mino.h-gq,goto.i-es,yamada.i-hy}@nhk.or.jp

概要

英語ニュースでは一般に一文内に同一の表現を繰り返し使うことを避け, 言い換えや省略によって簡潔な文にすることが推奨される. 一方で, 日本語ニュースではこうした簡潔性が英語ニュースほどは考慮されず, 一文内に同一の表現が複数回出てくることも珍しくない. この差分により, 日本語ニュースから機械翻訳で生成した文が英語ニュースのライティングスタイルに沿わないものになってしまうことがある. 本稿では, こうしたライティングスタイルを考慮した機械翻訳の実現に向けて, 時事通信社の日英ニュース記事対からテストセットを構築する. 日英ニュース記事の分析およびベースラインの機械翻訳モデルの性能評価を通じて, 本タスクの重要性および課題について議論する.

1　はじめに

ニューラルモデルの発展とともに機械翻訳の性能は大きく向上している. 機械翻訳の重要な応用先の一つがニュースであるが, 機械翻訳をニュース制作に活用する上での課題の一つとして, 言語間でのニュースのライティングスタイルの差分が挙げられる. 例えば放送用の英語ニュース原稿では一般に, 説明や主張を簡潔なものにするため, 同じ表現の繰り返しを避ける, 受動態よりも能動態が好まれる, 否定形よりも肯定形が好まれるといった特徴がある [1].

図 1 に [1] から引用した, 同じ表現の繰り返しを避ける例を示す. 例示した 2 つの文はいずれも「週末にハドソン百貨店で全米自動車労組がピケを行い, 双方が勝利を宣言している. 」という意味の文である. 1 つ目の文は「ハドソン百貨店」と「全米自動車労組」が複数回ずつ現れているが, 2 つ目の文はこれらの単語を一回ずつのみ記述している. 前者のように同一の名前や情報を繰り返す文 (REPETITIVE) は

(REPETITIVE)
Both Hudson's and the United Auto Workers Union are declaring victory after a weekend of U-A-W pickets at Hudson's Department Stores.
(TIGHTER)
Both sides are declaring victory after a weekend of picketing by the United Auto Workers Union at Hudson's Department Stores.

図 1　ライティングスタイルにそぐわない英文の例

冗長に感じられるため回避することが推奨され, より簡潔な文 (TIGHTER) が好ましい.

一方で, 日本語ニュースではこうした簡潔性が英語ニュースほどは考慮されず, 一文内に同一の表現が複数回出てくることも珍しくない. その結果, 日本語のニュース記事を機械翻訳で英訳すると英語ニュースのライティングスタイルにそぐわない文が生成されてしまうことがある. 図 2 に時事通信社の日本語ニュース文, 対応する英語ニュース文, 機械翻訳による英訳の例を示す. この例では日本語文内で「入学」という単語が 3 回出てくるが, 前述のライティングスタイルに則るとこれらを全て同一表現で訳出することは避けることが望ましい. 実際, 英語ニュース文では 1 つ目と 3 つ目の「入学」を 1 つにまとめ, 2 つ目の「入学」については “enter” ではなく “join” と表現している. しかし, 機械翻訳の出力では 2 つ目の「入学」については別表現を用いているものの, 1 つ目と 3 つ目の「入学」をいずれも同一に訳出してしまい, 冗長な文になってしまっている.

ニュースでは速報性が重要なため, 機械翻訳を制作ワークフローに導入するためには, なるべく人手の修正が少なくなるよう, 単に翻訳文の意味が合っているだけでなく上記のようなライティングスタイルの要件を満たす文を生成することが望ましい.

(SRC)
区によると, 同校がアルマーニの採用を決めた後, 入学予定者 7 人が教育方針や私立小への入学などを理由に入学を辞退した.
(REF)
After the school decided to recommend Armani items as desirable clothing, seven children dropped plans to enter the school, with parents citing disagreements with its education policy, decisions to join private schools or other reasons, according to the office of Chuo Ward, where the school is located.
(NMT)
According to the ward office, after the school decided to hire Armani, seven students who plan to enroll in the school declined to enroll in the school due to their educational policy and entrance to private primary schools.

図 2　ライティングスタイルにそぐわない翻訳の例

本稿ではニュースのライティングスタイルに沿った文の生成の実現に向けて, まず評価用のテストセットを構築する. 上述した「同一表現の繰り返しの抑制」を制御の対象として, 時事通信社の日英対訳記事から人手で事例を収集する. （以降, 本稿で単に「ライティングスタイル」と記述した際には, 同一表現の繰り返しの抑制のことを言及しているものとする. ）ニュースドメインの日英対訳コーパスで学習したベースライン機械翻訳モデルが, 上記のテストデータでどれくらいライティングスタイルに沿った出力を生成できているかを評価する. 日英ニュース記事の分析および評価実験を通じて, 本タスクの重要性および課題について議論する.

2　日英ニュース記事の分析

本研究では, 時事通信社の 2018 年の日英ニュース記事アーカイブを利用する. このアーカイブでは日英いずれの記事にも ID が付与されており, まずこの ID を照合して日英のニュース記事対を作る. 次に, 内山・井佐原の手法 [2] を用いて各記事対内で対訳文のペアを作る. 各日本語文を mecab [3] でトークナイズし, ストップワードを除き, 一文内でトークンが重複している文を列挙する. この中から, 日本語文を見て, （図 2 に示したような）同一の表現が繰り返されていて訳出時に言い換えや省略が起きうるものを手作業で 514 文収集した. これは「一文内でトークンが重複している文」のうちのおよそ 5, 6 文に 1 文程度に該当し, また, 記事単位でみるとおよそ 4 記事に 1 記事がこのような文を含んでおり, これはニュース制作に機械翻訳を応用する上では無視できない分量である. 抽出した 514 の対訳ペアの中には文レベルの自動対応付けがうまくいかず, 日本語側で繰り返されている表現が英文側でどのように訳出されているか判断できないものもあったため, この中で対応付けの精度が比較的良く, 対象の表現が英語側でどのように訳されているか判断できるものを人手で 100 文取り出し, 訳出の傾向を分析した.

省略　訳出時に省略が起きているもので単純なものとしては, “A and B” の形で書けるような, 文法構造上並列な名詞や動詞が挙げられる.

仮校舎と再開した校舎をテレビ会議システムで結び, 遠隔合同授業も行う予定だ.
Joint class activities using teleconference equipment are planned between the reopened and provisional schools.

また, もう 1 つ分かりやすいケースとして, 数量の後ろに単位として付く名詞が繰り返されている場合にもこれを省略する傾向にあった.

17 年 4 月現在, 福祉事務所がある 902 自治体の 56% に当たる 504 自治体が取り組んでいるが, 中学生の勉強をサポートする事業が大半となっている.
As of April 2017, 504 of 902 municipalities with welfare offices had implemented the program.

それ以外の事例は, 文法構造からは単純には判断できない複雑なものが多い.

民間調査機関インテージによると, 2017 年のアルコール度数 7〜9% の缶酎ハイ市場は 7 年前の 2.5 倍に拡大し, 缶酎ハイ市場全体の 5 割強を占めた.
But the share of products with 7-9 pct alcohol in the Japanese canned chuhai market grew 2.5-fold in seven years to stand at over 50 pct in 2017, according to private research firm Intage Inc.

これは日本語側で意味が重複しているため, 訳出し

てしまうと英語側が冗長になってしまうと考えられる例で, 2 つ目の「市場全体の 5 割」が「缶酎ハイ市場」であることは自明であるため省略されている.

言い換え　言い換えは積極的に用いられる傾向があり, 特に文内の近い位置で繰り返されているものについては言い換えが起きやすい.

> 昨年 3 月の声明は「戦争・軍事目的の科学研究を行わない」とする過去 2 回の声明を継承.
> In the March 2017 statement, the council pledged to follow its two previous documents highlighting its determination not to conduct scientific research for military purposes.

固有名詞の繰り返しについても, 言い換えが起きる. 以下の例では 2 つ目の「ソメイヨシノ」を “Someiyoshino” ではなく, “the cultivated variety” と言い換えている.

> クマノザクラの花は栽培品種のソメイヨシノ（染井吉野）に似た淡紅色などだが, 開花時期がソメイヨシノより早い.
> The petals of Kumanozakura are mainly a light shade of red, similar to those of “Someiyoshino,” and bloom earlier than those of the cultivated variety.

また, 表層が同じであってもニュアンスの異なる場合には訳出時に違う表現になる. 例として以下に示すような項の異なる述語が挙げられる.

> さらに, いずれのケースでもブレーキが作動する０．８秒前までに, 運転者に衝突回避操作を促す警報が作動することも認定の要件となっている.
> Another planned requirement for the ministry’s certification is that the equipment warns the driver at least 0.8 second before the brakes are activated, long enough for ordinary drivers to respond and apply brakes manually.

その他　日本語側で繰り返されている表現に対して, 英語側で必ずしも言い換え, あるいは省略が行われているわけではなく, 全て同一に訳出するものも存在する. これについては, 翻訳者が同一に訳してもさほど簡潔性を損なわないと判断した, その表現の訳語にバリエーションがなく同一に訳さざるを得なかったなどの理由が考えられる.

> 入居対象は市に定住する意思があり, 夫婦とも 35 歳以下か, 未就学児のみがいる夫婦. The city rents the apartments for up to five years to couples both aged 35 or less, or couples with only children of preschool age, who are willing to eventually settle in Hadano.

3　テストセットと評価方法

本稿では 2 節で収集した 100 組の対訳ペア（日本語入力文および参照訳）をテストセットとして用いる. テストセットの日本語文の中には 1 文内で複数種類の表現で繰り返しが起きているものもあり, 訳出の分類を合計すると省略が 45 事例, 言い換えが 39 事例, 全て同一に訳出していたものが 23 事例で合計 107 事例であった. また, テストセットの日本語文の平均文長は 60.52 文字であった.

次節で述べる実験では, 簡単のため参照訳と機械翻訳の出力を下記の 3 タイプに人手で分類し, これらがどれくらい一致しているかを評価する.

1. 対象の訳語が全て同一の場合は REPETITIVE
2. 対象の訳語が言い換えたり省略されている場合は TIGHTER
3. 対象の訳語が訳抜けや誤訳などによりうまく生成できていない場合は ERROR[1)]

参照訳の 107 事例については, 23 事例が REPETITIVE, 84 事例が TIGHTER となる. 本稿の実験では参照訳を正解として評価を行うが, 参照訳と機械翻訳の出力のタイプが一致していなかったからといって, それが必ずしも誤りであるとは限らない. ライティングスタイルは属人的なものであり, 言い換え・省略を行うか行わないかは作業者によって異なりうるからである. 現在, 複数人の翻訳者によるテストセットの日本語文の翻訳を進めており, 結果を収集してそのばらつきを調査する予定である.

4　実験

ベースラインの実験として, 日英ニュースドメインの対訳データで学習した機械翻訳モデルがどれほどライティングスタイルに沿った出力を生成できるかを評価する.

1)　2 項目の省略と 3 項目の訳抜けを真に見分けることは難しいが, ここでは原言語文の後半の節が丸ごと訳出できていないなど, 欠落の範囲が大きなものを訳抜けとして判定する.

表 1　実験結果

		Baseline			
		REPETITIVE	TIGHTER	ERROR	計
Gold	REPETITIVE	13	3	7	23
	TIGHTER	40	28	16	84
		53	31	23	107

読売新聞 (Yomiuri_Editorial)[2] の 2007 年から 2017 年の日英対訳記事コーパスから類似度 0.4 以上の対訳ペアを抽出し, このうち 596,679 ペアを学習データ, 1,583 ペアを開発データとして用いる. いずれの言語もトークナイズに sentencepiece [4] を用い, 語彙サイズを 8000 とした. 翻訳モデルは Transformer [5] とし, fairseq [6] で実装した. optimizer は Adam [7] を用い, 学習率は 0.001 とした. 翻訳時のビームサーチはビーム幅を 5 とした.

表 1 に実験結果を示す. 全体で参照訳と一致していたのは 107 事例中 41 事例にとどまった. TIGHTER について一致したのは 28 事例で, さらにそのうち繰り返すか省略するかについても一致していたのは 19 事例であった. ベースラインモデルは全体の半数程度が REPETITIVE な出力となっており, 言い換えや省略の制御機能を学習データからうまく獲得する工夫が必要であることを示している. また, ERROR の数も 1/5 程度となっており, 無視できない問題であることが分かる. ERROR の内訳はほとんどが訳抜けであった. テストデータの中には比較的長い文が多く, これらをうまく処理できていない. 日本語ニュースは他ドメインと比べて文長が長い傾向があるが, 文長が長くて類似度も高い訓練事例は多くないため, 文長にロバストな翻訳性能をいかにして獲得するかも課題である.

5　関連研究

機械翻訳における訳語の統一については多く研究がなされてきた一方で [8, 9, 10], 訳語の言い換えについての研究が限られている. 同一の単語はなるべく同一に訳したほうが可読性は高まるが, あまりに同一表現を使いすぎるとかえって読み物としての質が下がってしまう恐れもあるため, 一概にどちらにすべきとは言えない. こうした背景を受けて, Guillou [11] は実世界の色々なドメインの翻訳において, どのような単語が言い換えられやすい（あるいは統一されやすい）傾向にあるのかを統計的に調査した. 単語の頻度や品詞, 文書のドメインによって傾向は異なるものの, 訳語を適宜言い換えることの重要性を提起している. Guillou [11] の研究は分析を主としたものだが, 本研究はニュースドメインにおける訳語の言い換えを実際のタスクに落とし込むことをねらいとしている.

翻訳のスタイルの制御についても, 近年注目を集めている. Wang ら [12] は翻訳者の翻訳スタイルをニューラルモデルに学習させる研究を行った. 森下ら [13] はミニバッチに一文書を丸ごと入れることで文脈を考慮しながら学習・推論を行う手法を提案し, その効果の 1 つとして,「です・ます」調や「だ・である」調など, 翻訳対象のドメインのライティングスタイルに沿った出力が得られたことを報告している. 本研究では翻訳スタイルの 1 つとして, ニュースのライティングスタイルの一項目を対象に取り上げ, この制御を評価することを目的としている.

6　おわりに

本稿では日英間のニューススタイルの差分に着目し, このギャップを埋めるための翻訳制御の実現に向けて, テストセットを構築した. 日英のニュース記事対についてそのライティングスタイルの差分を分析し, 本タスクの重要性とベースラインモデルの課題について議論した.

今後の予定として, 本稿で実施した評価方法についての検討を進める. 言い換えと省略を 1 つに混ぜて評価を行なったが, どのような場面でどちらかが求められているかについても調査を行う. また, 現在複数名の翻訳者により, 収集した 514 の日本語文の翻訳を進めており, その結果を受けてライティングスタイルのばらつきなどについての分析も進める予定である. どのような誤りがニュース制作者にとってインパクトの大きなものかについても定性的に分析したい.

さらに、実際にこうしたライティングスタイルを文生成に組み込むための効果的な手法についても検討する. 本稿の実験では翻訳モデルの出力を評価したが, 本タスクは翻訳モデルの前後の処理でも効果的な制御を行える可能性がある. 例えば, 日本語側を pre-edit する [14, 15], あるいは, 英語側を post-edit するなどの手法も考えられる.

2)　https://www.nichigai.co.jp/dcs/index5.html

謝辞

本研究成果は, 国立研究開発法人情報通信研究機構の委託研究（課題 225）により得られたものです. また, データをご提供頂きました株式会社時事通信社の朝賀英裕氏・川上貴之氏に厚く御礼申し上げます.

参考文献

[1] Robert A. Papper. **BROADCAST NEWS AND WRITING STYLEBOOK**. Routeledge, seventh edition, 2021.

[2] Masao Utiyama and Hitoshi Isahara. A Japanese-English patent parallel corpus. In **Proceedings of Machine Translation Summit XI: Papers**, Copenhagen, Denmark, September 10-14 2007.

[3] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 230–237, Barcelona, Spain, jul 2004. Association for Computational Linguistics.

[4] Taku Kudo and John Richardson. 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, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[6] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, **3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings**, 2015.

[8] Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. Modeling coherence for neural machine translation with dynamic and topic caches. In **Proceedings of the 27th International Conference on Computational Linguistics**, pp. 596–606, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[9] Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. Learning to remember translation history with a continuous cache. **Transactions of the Association for Computational Linguistics**, Vol. 6, pp. 407–420, 2018.

[10] Xinglin Lyu, Junhui Li, Zhengxian Gong, and Min Zhang. Encouraging lexical translation consistency for document-level neural machine translation. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 3265–3277, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[11] Liane Guillou. Analysing lexical consistency in translation. In **Proceedings of the Workshop on Discourse in Machine Translation**, pp. 10–18, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[12] Yue Wang, Cuong Hoang, and Marcello Federico. Towards modeling the style of translators in neural machine translation. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1193–1199, Online, June 2021. Association for Computational Linguistics.

[13] Makoto Morishita, Jun Suzuki, Tomoharu Iwata, and Masaaki Nagata. Context-aware neural machine translation with mini-batch embedding. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 2513–2521, Online, April 2021. Association for Computational Linguistics.

[14] Rei Miyata, Anthony Hartley, Cécile Paris, Midori Tatsumi, and Kyo Kageura. Japanese controlled language rules to improve machine translatability of municipal documents. In **Proceedings of Machine Translation Summit XV: Papers**, Miami, USA, October 30 – November 3 2015.

[15] Yusuke Hiraoka and Masaru Yamada. Pre-editing plus neural machine translation for subtitling: Effective pre-editing rules for subtitling of TED talks. In **Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks**, pp. 64–72, Dublin, Ireland, August 2019. European Association for Machine Translation.