字幕機械翻訳における自動訳抜け検出の試みとその分析

石川隆太 [1]　須藤克仁 [1]　松島朝子 [2]　中村哲 [1]
[1] 奈良先端科学技術大学院大学　[2] 日本映像翻訳アカデミー株式会社
{ishikawa.ryuta.il7,sudoh,s-nakamura}@is.naist.jp
subit@jvtacademy.com

概要

本研究では、英語から日本語への字幕翻訳を題材として、機械翻訳結果における訳抜けの自動検出を試み、その結果の分析を行う。字幕翻訳では字幕表示の制約により省略や言い換えを通じて訳出が短くなる傾向にあり、そうした事例を学習した機械翻訳は訳抜けをしやすく、自動訳抜け検出により翻訳後編集の効率向上が期待される。映像字幕データに対する実験では、自動訳抜け検出は一定の効果が認められるものの、字幕翻訳の特性に起因する訳抜けの過剰検出が目立ち、省略と訳抜けのトレードオフ解消の難しさが示唆される結果が得られた。

1　はじめに

字幕翻訳は映像素材に付与された原言語の台詞や挿入テキストを目的言語に翻訳するタスクであり、その対象は映画に留まらずドラマやドキュメンタリーのような各種映像作品に及ぶ。字幕翻訳では、素材に適し、かつ映像への重畳表示を踏まえた翻訳を行うための様々な戦略が存在する [1]。著者らは字幕翻訳を対象とする機械翻訳について、英語から日本語への翻訳を対象に検討を行っている。字幕翻訳のデータを用いたニューラル機械翻訳 (NMT) モデルの学習、様々な素材に対する機械翻訳結果と日本語字幕と比較しての自動評価、映像翻訳者による評価・分析を通じて、字幕機械翻訳における様々な課題が明らかになってきており、特に訳抜け (omission あるいは under-translation) が顕著である。機械翻訳と翻訳後編集 (post-editing; PE) による翻訳ワークフローにおいて、訳抜けは翻訳の効率に大きな影響を及ぼすため、その抑制が強く求められる。しかしながら、NMT は仕組み上訳抜けを完全に防ぐことが難しい上、限られた文字数で情報を伝えるための意訳や省略がしばしば行われる字幕翻訳の事例を学習した NMT では訳抜けがより生じやすくなっていると考えられる。

そこで本研究では、翻訳後編集の際に訳抜けが疑われる箇所を強調表示する等のユーザインタフェースを提供することを念頭に、字幕機械翻訳に対して既存の自動訳抜け検出手法を適用し、その評価と結果の分析を行った。訳抜け検出の Precision は平均で 20%強、Recall は平均で 40%程度で課題が残る結果となったが、分析により訳抜けの検出漏れや過剰検出で特徴的ないくつかのパターンが明らかになった。本研究で得られた結果は、原文の情報をできるだけ忠実に訳出するタイプの翻訳と比べ、字幕機械翻訳における望ましい省略と望ましくない訳抜けのトレードオフ解消の難しさを示唆するものである。

2　関連研究

自動訳抜け検出の既存研究 [2] では NMT モデルのみを利用した手法を提案している。次節で述べる通り、本研究ではこの手法を用いて字幕機械翻訳における訳抜け検出を試みる。それ以前の研究では訳抜けや余分な訳の評価が試みられていた [3, 4] が、そのために参照訳を要するという点で問題が異なる。

自動訳抜け検出と関係する課題として、単語レベルの機械翻訳の品質推定 [5] が挙げられる。機械翻訳の品質推定は機械翻訳結果中の誤訳の程度や箇所を参照訳を用いずに推定するタスクであり、特に単語レベルでの品質推定では、訳文中の各単語に対する正誤判定と、訳文中で抜けが生じている箇所の予測を行っている。近年の共通タスク [6] は MQM [7] に基づく訳文中の誤りスパンを予測するもので本研究で扱う問題と一見類似するが、品質推定タスクでは訳文中の誤り箇所を同定するのに対し、自動訳抜け検出では原文中で訳出されなかった箇所を同定するという違いがあり、単語レベル品質推定の既存データやモデルはそのままでは適用できない。

また、字幕翻訳に関しては、TED Talks の字幕を利用した WIT^3 [8] や MuST-C [9] が音声翻訳の研究

【原文と訳文】
原文 X = *Please call me when you arrive.*
訳文 Y = 電話してください

【部分単語列に対する翻訳スコアの計算】
Score (Y | *Please call me when you arrive.*) = 0.42
Score (Y | ~~*Please*~~ *call me when you arrive.*) = 0.29
Score (Y | *Please call* ~~*me*~~ *when you arrive.*) = 0.41
Score (Y | *Please call me* ~~*when you arrive*~~.) = **0.67**

【訳抜け箇所の同定】
Please call me <omit>*when you arrive*</omit>.

図 1　自動訳抜け検出 [2] の概略

で広く用いられてきた他、多くの言語をカバーする OpenSubtitles [10] や、日英の言語対に特化した JESC [11] も公開されている。これらのデータを用いた英日翻訳の実験結果では BLEU [12] で 10 程度のスコアであることが多く、訳出の多様性による評価の難しさは考慮するとしても、機械翻訳の対象として依然難しいことが窺われる。本研究は、様々な分野の素材を含む独自の字幕翻訳の英日対訳データを利用し、訳抜けの問題に焦点を当てて自動検出を試み、問題点を明らかにすることを狙うものである。

3　自動訳抜け検出

Vamvas ら [2] は統語構造上の部分木をランダムに削除した入力文を用いた自動訳抜け検出手法を提案している。この手法では Contrastive Conditioning [13] と呼ばれる、機械翻訳の入力側に小さな変更を加え、所与の翻訳出力を teacher forcing によって得た場合に生じるスコアの違いを翻訳誤りの推定に用いる手法を応用している。

Vamvas らの自動訳抜け検出手法の概略を図 1 に示す。部分単語列は依存構造木中の部分木を削除することによって得られる。もし元の原文 X よりもその部分単語列 X' を用いたほうが高い翻訳スコアが得られるのであれば、X' を得る際に削除した部分木に相当する箇所 X^{del} （図 1 中で取り消し線が付された箇所）は訳文 Y への翻訳において不要である、すなわち、訳文 Y には該当する内容が訳出されていない訳抜け箇所であると判断できる。本手法は直感的であるとともに、機械翻訳のスコアのみで判断を行うため、機械翻訳の内部機構に依存せず適用可能である点で有益である。したがって、この手法を本研究の訳抜け検出に利用する。

表 1　評価・分析に用いた字幕数、NMT による翻訳結果で訳抜けと判定された字幕数、および参照訳に対する BLEU と長さ比 (LR)

ジャンル	字幕数	訳抜け判定数 重度	訳抜け判定数 軽度	BLEU (LR)
ドラマ	46	4	10	12.9 (0.88)
リアリティショー	50	7	6	9.7 (1.15)
ビジネス講演	270	45	36	9.7 (0.74)
ドキュメンタリー	218	28	28	8.8 (0.88)

4　実験

Vamvas らの訳抜け検出手法によって字幕機械翻訳における訳抜けがどの程度検出できるかを調査するため、実際の字幕機械翻訳のデータを用いた以下の訳抜け検出実験を行った。

4.1　実験設定

字幕機械翻訳には fairseq[1] [14] により構築された Transformer による NMT を用いた。学習データは字幕翻訳の英日対訳データ約 550 万対と、その他 WWW からクローリングを通じて抽出した英日対訳データ約 1,500 万対である。

自動訳抜け検出には Vamvas らの実装[2]を用いた。なお、当該実装では与えた原言語字幕中の先頭の文のみを対象に削除する部分木を選択しており、しばしば 2 文以上の短い文が含まれる字幕のデータの処理に問題があった。そこで、本実験では 2 文以上の入力に対してすべての文から削除する部分木が選択されるように修正を加えて実験に利用した。依存構造解析には当該実装の標準設定に基づき Stanza[3] [15] の英語モデルを用いた。また、翻訳スコアの計算を行うための NMT モデルは、原論文で用いられている Huggingface の `mbart-large-50-one-to-many-mmt`[4] を、字幕翻訳の英日対訳データ約 550 万対を用いてファインチューニングしたものを用いた。

評価および分析にはファインチューニングに用いたものとは別の映像翻訳字幕の英日対訳データを用いた。映像素材のジャンルの違いによる傾向の差を分析するため、ドラマ、リアリティショー、ビジネス系講演、ドキュメンタリーの 4 ジャンルの素材

1) https://github.com/facebookresearch/fairseq
2) https://github.com/ZurichNLP/coverage-contrastive-conditioning
3) https://github.com/stanfordnlp/stanza
4) https://huggingface.co/facebook/mbart-large-50-one-to-many-mmt

表 2　自動訳抜け検出の正解・検出漏れ・過剰検出の数および Precision (P) 、Recall (R)

ジャンル	正解	漏れ	過剰	P (%)	R (%)
ドラマ	6	6	16	27.3	42.9
リアリティショー	5	6	22	18.5	38.5
ビジネス講演	27	29	110	19.7	33.3
ドキュメンタリー	24	14	89	21.2	42.9

を用いた。なお、ファインチューニングに用いた対訳データにもこれらのジャンルのものは含まれていた。ジャンル毎の字幕数や、字幕翻訳者によって判定された NMT 結果の訳抜けを含む字幕数、および参照訳に対する BLEU の情報を表 1 に示す。ジャンルにより訳抜け数や BLEU、NMT 訳と参照訳の長さ比には違いがあり、特に長さ比についてはリアリティショージャンルでは参照訳より 15%程度長く、ビジネス講演ジャンルでは 26%短かった。また、BLEU の値は 10 前後であり、2 節でも述べた従来研究における BLEU の値と同等と言える。

4.2　実験結果

表 2 に自動訳抜け検出の結果を示す。評価は字幕 1 枚を単位として正解、検出漏れ、過剰検出[5]に分類する形で行った。自動訳抜け検出された箇所が字幕翻訳者による訳抜け判定箇所と一致した場合を正解とし、訳抜け箇所が存在するにもかかわらず一切訳抜けが検出されなかった場合は検出漏れ、その他無関係な箇所での検出は過剰検出として計上した[6]。

実験の結果、過剰検出が明らかに多く、Precision（=正解字幕数/訳抜け自動検出字幕数）は 20%前後であり、全体として訳抜けを過剰に検出していることが分かった。一方、Recall（=正解字幕数/訳抜け判定字幕数）は全体で 40%をやや下回る程度で、ビジネス講演ジャンルでは他ジャンルと比べて見落としが多いことが分かった。表 1 に示した訳抜け数や長さ比 (LR) からも分かる通り、ビジネス講演ジャンルでは訳抜けの指摘数が多く、自動訳抜け検出でも他ジャンルに比べて性能が劣ることが明らかになった。より詳細な分析については次節で述べる。

5）　本研究では字幕翻訳者によって問題のある訳抜けを判定していることから、実際には訳出されているものを自動訳抜け検出が誤って検出した誤検出に加え、字幕翻訳者が問題視しなかった訳抜けを検出した場合も含める形で、誤検出ではなく**過剰検出**と呼ぶこととする。

6）　つまり、字幕翻訳者による訳抜け指摘と異なる箇所で訳抜けが自動検出された場合には検出漏れではなく過剰検出と計上されている。したがって、表 2 において正解数と検出漏れ数の和は表 1 の訳抜け数と一致するとは限らない。

5　分析

訳抜け検出の性能についてより詳細に調査するため、訳抜け検出結果の詳細な分析を行った。分析により判明した訳抜け検出の漏れ・誤りの傾向について以下詳しく述べる。

5.1　検出漏れ

訳抜けの検出漏れは副詞節として働く従属節や、目的語として that 節を取っている主節においてよく見られた。これらは文の意味上の焦点からやや外れていることもあり、字幕翻訳においては文脈から内容が明らかな場合に省略されたり、短い表現への意訳や言い換えが行われることがある。そのため、字幕翻訳データによって学習された NMT モデルはこれらの直接的な訳出に抑制的になり、訳抜けに繋がると考えられる。同様に、字幕翻訳用 NMT モデルを訳抜け検出に用いた場合、そうした表現の有無がスコアに大きな変化をもたらさないと考えられる。

他にも、特に検出漏れの多かったビジネス講演ジャンルにおいては、主節であっても代名詞が主語や目的語に含まれている箇所での訳抜けの検出漏れが多く見られた。主節の訳抜けも前段と同様の字幕翻訳の特性に起因するものと考えられるが、内容理解の妨げとなり、訳質の観点ではより深刻な問題である。そうした訳抜け検出が強く期待される箇所での検出性能に課題を残す結果であったと言える。

5.2　過剰検出

訳抜けの過剰検出については多くの要素で起こっていたため、特徴的な 5 種について述べる。

副詞的役割を持つ要素　訳抜けの過剰検出において顕著だったものは、副詞・副詞句・副詞節である。特に程度を表す副詞である very や so 等は字幕翻訳においては直接的に訳出されない傾向が強く、形容詞や選択の工夫により文字数の削減に寄与しているものと考えられる。そのため字幕翻訳データによって学習された NMT モデルを用いた機械翻訳においては訳出されづらい。一方で今回の分析を通じ、自動訳抜け検出は very や so の直接的訳出がないことの検出力が高いことが判明した。これは、字幕翻訳用の NMT モデルは翻訳時に直接的な訳出をもたらすほどの強い変化はもたらさないが、Contrastive Conditioning による翻訳スコアの変化には寄与しうることを示唆している。なお、その他の一般の副詞

や、副詞句・副詞節については、very や so ほど強い傾向は見られないものの、他の種類の要素と比べて訳抜けとして検出される傾向が強かった。

弱い関係の接続詞・感動詞・「つなぎ」の表現　次に顕著だった要素として、so や then のような弱い関係を表す接続詞、wow や oh 等の感動詞、you know 等の「つなぎ」の表現が挙げられる。これらは会話や台詞等で頻繁に用いられるが、程度を表す副詞と同様、字幕翻訳において直接的に訳されることは少ない。前段の結果も含め、Contrastive Conditioning はこうした要素を敏感に検出していると言える。

誤訳された要素　本実験では各字幕をそれぞれ独立に翻訳しており周辺文脈を参照していないこと、また素材毎の用語対訳集等も与えていないことから、翻訳に必要な情報の不足による誤訳が多く生じていた。そのように誤訳された原文の箇所が訳抜けとして検出された事例、特にハイコンテクストな訳出が求められる箇所、稀少語や固有名詞等で多く見られた。特に周辺文脈の問題に関しては、映像字幕において文字数の制約により 1 文を複数の字幕ページに分割する傾向も考慮し、NMT、自動訳抜け検出の双方で周辺の字幕ページを文脈情報として参照することが有益であろうと考えられる。

並列・反復表現において複数回出現する要素　出現数は多くなかったものの注目すべき事例として、並列や反復表現において同一の単語や句が用いられていた場合に自動訳抜け検出で訳抜けでない箇所を検出するものがあった。Contrastive Conditioning の性質上、並列・反復表現で複数回出現する単語の一方を削除しても、残った単語から NMT が補完可能であれば訳抜けと検出される可能性がある。この点は本実験における訳抜け検出手法の本質的な問題であり、異なるアプローチが必要であろう。

字幕中の余剰文字列　複数の出演者が存在する素材においては、字幕に話者を表す文字列が挿入されることがあり、字幕の内容本体と区別されずに機械翻訳を行うとそれが翻訳に悪影響を与えてしまう場合がある。本実験においても、ドラマおよびリアリティーショーの素材で各字幕の先頭に話者名を表す文字列が挿入されている箇所では、翻訳時に話者情報が削除された場合その箇所が訳抜けしていると検出されるという問題が生じていた。話者情報は字幕に付随する情報として重要でもあるため、字幕の内容本体と区別した上での利用が有効と期待される。

6　議論

ここまでで示した通り、実験における自動訳抜け検出の性能はまだ十分とは言えず、様々な要素で検出漏れや過剰検出が生じていた。

字幕翻訳においては限られた時間で情報を伝達しなければならないため、文脈も踏まえた情報の取捨選択や、表現の言い換えや意訳等を通じた文字数の削減が行われる（原音 1 秒あたり日本語 4 文字、1 行あたりの文字数は 14〜16 文字、等の目安が存在する [1]）。今回字幕翻訳に用いた NMT モデルはそうした高度なプロセスを経た字幕翻訳の事例を学習に用いているものの、実際に翻訳を行う際には学習事例の統計的な傾向をもとに省略や言い換えを行うに過ぎない。その結果が時に有用であることもあるが、字幕翻訳者と同様の思考プロセスを経たものではなく、適切な訳出に繋がらず訳抜けや冗長な訳出をもたらすことは否定できない。

さらに、訳抜け検出に使う NMT モデルも翻訳に用いたモデルと本質的には同様の問題を持つ。Contrastive Conditioning による効果は一部認められるものの、原文の要素の訳出要否の判断に原文以外の情報が利用できないことによる限界は存在する。この点は単語レベルの機械翻訳品質推定 [5] と同様であるが、一般的な翻訳と比べて字幕翻訳では情報の取捨選択が与える影響が大きく、意図的に訳出を抑制する省略と、意図せず起こる訳抜けの区別の重要性の面でより難しい問題であると言える。

7　おわりに

本研究では、字幕翻訳において機械翻訳と翻訳後編集を活用するワークフローの効率化を見据え自動訳抜け検出手法を適用し、その結果の分析を行った。Vamvas らの Contrastive Conditioning に基づく自動訳抜け検出により 40% 程度の Recall で訳抜けの検出ができた一方で Precision は 20%程度に留まり、訳抜けの過剰な検出が目立つ結果となった。これは依然課題が残る結果と言えるが、訳抜けの検出漏れや過剰検出は省略や意訳の多い字幕翻訳から学習された NMT モデルにとってはトレードオフの関係となっていると考えられ、この解決には更なる工夫、例えば文脈情報を活用した NMT および訳抜け検出等が求められるだろう。

参考文献

[1] 日本映像翻訳アカデミー. 字幕翻訳とは何か　1 枚の字幕に込められた技能と理論. 2018.

[2] Jannis Vamvas and Rico Sennrich. As little as possible, as much as necessary: Detecting over- and undertranslations with contrastive conditioning. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 490–500, Dublin, Ireland, May 2022.

[3] Jing Yang, Biao Zhang, Yue Qin, Xiangwen Zhang, Qian Lin, and Jinsong Su. Otem&utem: Over- and undertranslation evaluation metric for nmt. In **Natural Language Processing and Chinese Computing**, pp. 291–302, Cham, 2018.

[4] Xiang Kong, Zhaopeng Tu, Shuming Shi, Eduard Hovy, and Tong Zhang. Neural machine translation with adequacy-oriented learning. In **Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2019)**, pp. 6618–6625, Jul. 2019.

[5] Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. Findings of the WMT 2019 shared tasks on quality estimation. In **Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)**, pp. 1–10, Florence, Italy, August 2019. Association for Computational Linguistics.

[6] Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. Findings of the WMT 2022 shared task on quality estimation. In **Proceedings of the Seventh Conference on Machine Translation (WMT)**, pp. 69–99, Abu Dhabi, United Arab Emirates (Hybrid), December 2022.

[7] Arle Lommel, Hans. Uszkoreit, and Aljoscha Burchardt. Multidimensional Quality Metrics (MQM) : A Framework for Declaring and Describing Translation Quality Metrics. **Tradumàtica**, No. 12, pp. 455–463, 2014.

[8] Mauro Cettolo, Christian Girardi, and Marcello Federico. WIT3: Web inventory of transcribed and translated talks. In **Proceedings of the 16th Annual Conference of the European Association for Machine Translation**, pp. 261–268, Trento, Italy, May 28–30 2012.

[9] Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. MuST-C: a Multilingual Speech Translation Corpus. 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. 2012–2017, Minneapolis, Minnesota, June 2019.

[10] Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**, pp. 923–929, Portorož, Slovenia, May 2016.

[11] Reid Pryzant, Youngjoo Chung, Dan Jurafsky, and Denny Britz. JESC: Japanese-English subtitle corpus. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Miyazaki, Japan, May 2018.

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, Philadelphia, Pennsylvania, USA, July 2002.

[13] Jannis Vamvas and Rico Sennrich. Contrastive conditioning for assessing disambiguation in MT: A case study of distilled bias. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 10246–10265, Online and Punta Cana, Dominican Republic, November 2021.

[14] 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.

[15] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**, pp. 101–108, Online, July 2020.