言語処理学会第30回年次大会　発表論文集（2024年3月）
*発表論文集発行が3月以降で提出締切に間に合わないため、手書きにて学会名記入

同時通訳・同時翻訳のための語順同期性評価

蒔苗茉那　須藤克仁　中村哲
奈良先端科学技術大学院大学
{makinae.mana.mh2, sudoh, s-nakamura}@is.naist.jp

概要

英日のような文法が大きく異なる言語ペアにて、同時通訳者はリアルタイムで訳出を行うために、原発話の語順を可能な限り維持して遅延を最小化しつつ品質を維持しているとされている。これは原発話の語順と同期した出力が望ましいことを意味し、機械による同時通訳や同時機械翻訳もこのような訳出スタイルを模倣することが発展の鍵となる。そこで本研究では順位相関係数と多言語 BERT を使用して、語順同期性に着目した同時通訳 (SI) と同時翻訳 (SiMT) の自動評価指標を提案する。NAIST-SIC-Aligned と JNPC コーパスを用いた実験の結果、我々の評価指標が原発話と目的発話の間の語順同期の程度を計るのに有効であることが示された。

1　はじめに

SI は原発話の終了前に訳出を開始する特徴をもち、高品質かつ低遅延な訳出を行うことが目標である。その中で、特に英語と日本語のような異なる語順の言語対で高品質と低遅延の両立を図るためには、可能な限り原発話と同期させることが重要である。具体的に、同時通訳者は良質かつ最小の遅延を達成するために、できるだけ原発話の語順を保持した状態で訳出を行うこともあり、この戦略はファーストイン・ファーストアウト（FIFO）戦略と呼ばれている [1, 2, 3]。近年 SI に対して計算機を用いた同時機械翻訳（SiMT）への研究関心が高まっているものの、入手可能な SI コーパスのサイズが小さいこともあり、ほとんどの研究 [4, 5] の学習で用いられているデータの大半は同時通訳者が行っているとされている FIFO 戦略が反映されていない通常の翻訳データである。つまり SiMT は FIFO の戦略を通じた品質と遅延を両立させた訳出ではなく、機械翻訳と同じような語順の並び替えの距離が長い訳出を用いた学習が行われている。表 1 は、翻訳と SI の間の英日の語順差の例を示している。SI について、原発話では「every year」は文末に来ているが、「every year」にあたる日本語訳は文の中ほどに来ている。これは、遅延を減らすために原発話の途中で既に訳出が始まっていることを示している。一方、翻訳では「every year」にあたる日本語訳が文頭に来ている。これは、訳出の開始が原発話の終了まで待たなければならないことを示し、これは遅延の増加につながることが示唆される。

そこで本研究では、SI と SiMT の 原発話と目的発話の間の語順同期性を計る評価指標を構築する。提案手法はこれまでの自動機械翻訳評価の研究から影響を受けている。RIBES [6] の順位相関係数の考えを用いて語順同期性を計算する。順位相関係数を計算する際に必要なアライメント情報の取得について、従来手法は表層ベースであるが、提案手法は BERTScore [7] を用いた意味的類似度から取得する。基本的に BERTScore はモノリンガルであるが、多言語 BERT を用れば容易にクロスリンガルに拡張できる [8]。提案手法の有用性を英日 SI コーパス [9, 10] を用いて検証したところ、特に長い原発話において SI は翻訳よりも語順同期性が高いことが示された。

2　関連研究

2.1　同時通訳の人手評価

人間による SI の評価指標として主に：NAATI（National Accreditation Authority For Translators and Interpreters）Metrics [11] と EU Metrics [12] がある。これらの評価指標は文単位の訳出結果ではなく、通訳者のパフォーマンスを総合的に評価している。それに対して、本研究は文単位で SI の語順同期性を評価する。

2.2　機械翻訳の自動評価

BLEU [13] は機械翻訳で標準的に使用されている自動品質評価であり、SiMT でも活用されている

原発話書き起こし	(1) Out of seven / (2) large public corporations / (3) commit / (4) frauds / (5) **every year**.
同時通訳 書き起こし	上場している企業の / 7 社に 1 社は / 毎年 / 不正行為を / しています。 [(2) large public corporation / (1) out of seven / (5) **every year** / (4) frauds / (3) commit]
字幕（翻訳）	毎年 / 大企業の / 7 社に 1 社が / 不正行為を / 働いています。 [(5) **every year** / (2) large public corporation / (1) out of seven / (4) frauds / (3) commit]

表 1　英日における語順差の例

[5]。 RIBES は順位相関係数を用いて単語の順序の違いを考慮した自動評価である [6]。近年の多次元空間への埋め込みを用いた自動評価は、従来の表層一致評価と大きく異なり、単語の意味に基づいた評価を行っている。BERTscore [7] は、hypothesis と reference 間の明示的なトークンアラインメントからそれぞれのトークンの類似性を用いた評価を行っている。COMET [14] は、マルチリンガルな事前学習モデルを活用した文レベルの埋め込みを使用しており、COMET-QE [15] は参照訳を用いることなく、直接 hypothesis と reference からの評価を可能としている。提案手法は、語順同期性を計るために、RIBES、BERTScore、および COMET-QE から着想を得ている。

2.3　遅延評価

従来の SiMT の研究では、Average Lagging [5] およびその派生形 [16, 17] を用いた遅延評価が主流である。これらの手法はタイムスタンプを用いて遅延を計算するため、入力と出力の意味的な対応は考慮していない。SI の研究では、Ear Voice Span(EVS) [18] を使用しており、これは原発話と対応する同時通訳の訳出間の時間差に基づいた遅延評価である。本研究でも、遅延を削減する技術に注目しているが、時間差を表すタイムスタンプの代わりに、SI と SiMT の遅延に影響するであろう単語の語順同期性の数量化を目的としている。テキストレベルの情報を見ることができれば、どの訳出が遅延を発生させているか原因を特定でき、その結果を用いて遅延削減のためにどのように訳出を工夫すれば良いか示唆を得ることが可能だからである。

3　前提条件

本研究では、入力と出力の間の語順同期性を計るために、これまでの自動機械翻訳評価指標で用いられた以下の技術を使用している。

3.1　BERTScore によるアライメント取得

BERTScore は、貪欲法によるトークンアラインメントを使用して、hypothesis トークン $h_1, ..., h_n$ と reference トークン $r_1, ..., r_n$ に対して $a_i = \arg\max_j \cos(\boldsymbol{h}_i, \boldsymbol{r}_j)$ をとる。ここでは、a_i は h_i に対応する参照トークンのインデックスであり、h_i と r_i は埋め込みベクトルである。本研究では、多言語 BERT を用いて元の単一言語 BERTScore をクロスリンガルに適用する。このようなアライメント問題に対して、バイリンガルな単語アラインメント [19, 20, 21] が一般的である。しかし同時通訳の訳出結果は翻訳と異なり表層レベルで一対一対応していないもの（要約、省略などを含む [1]）が大部分を占めるため、予備実験にて十分な結果を得ることができなかった。

3.2　RIBES による順位相関係数

[6] は、Kendall の順位相関係数 τ を使用して hypothesis と reference の語順同期性を計る。RIBES では単語の表層一致で自動的にアライメントを取得するが、提案手法では BERTScore から得られる意味的一致を用いて、対応が取れたトークンのインデックスを使用して順位相関係数を計算する。

4　提案手法

本研究では、SI および SiMT を対象に言語横断のトークン対応付けを使用した語順同期性を計る評価指標を提案する。例えば、英語の入力が「I ate apples yesterday.」であり、それに対応する日本語の出力が「私は (I) / 昨日 (yesterday) / りんごを (apple) / 食べました。(ate)」であると仮定する。ここで、スラッシュは英語の単語と日本語の文節の境界を示している。なお、これは単純化された例であり、実際のデータではサブワードでトークン化されている。アラインメントが取れた場合、翻訳で発生した単語の並べ替えを表す整数のリスト [1, 4, 3, 2] が得

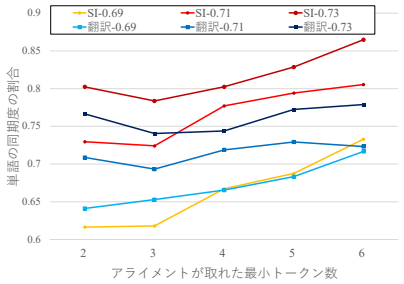

図 1　アライメント閾値別での提案手法のスコア推移

られる。本研究では、並べ替え距離を考慮するために Kendall's τ の代わりに Spearman's ρ の順位相関係数を使用する。上記の例では、ρ は 0.2 であり、ほとんど順位相関がないことを示している。前述にあるような同時通訳の特徴から、ここでのアラインメントの多くはノイズを含んでいる。これに対処するために、信頼性のあるアラインメントか否かを識別するために 2 つのヒューリスティックを適用する。まず、英語側にて機能語に対するアラインメントを語順同期性を計算する対象から外す。これは機能語は異なる言語間で常に明確な単語レベルの対応関係を持っているとは限らないからである。これについて、SpaCy の機能語リストを使用し、品詞タグに基づいて機能語を識別する。次に、類似度がある閾値 θ を下回るアライメントは信頼性に欠けるとしてそれらアラインメントも計算の対象から除外する。

5　実験

5.1　実験設定

提案手法の有効性を実際の SI および翻訳にて検証するために、以下の 2 つの実験を行った。

第 1 の実験では、翻訳と SI の語順同期性の程度を比較した。ここでは NAIST-SIC-Aligned [9] を使用した。これは TED Talks の英日同時通訳を集めたものである。本実験では 15 年以上の経験を持つ同時通訳者による 212 トークのみを抽出し、高品質な通訳に焦点を当てて検証している。経験の浅い同時通訳者による SI は誤訳などが多く、順位相関を計算するためのアライメントをとることが難しかった。これら SI の分析は今後の課題の一つとする。

第 2 の実験では、提案手法と人手評価との相関関係を調査した。ここでは Japan National Press Club（JNPC） Interpreting Corpus [10] から 1 つのトークを使用した。SI の人手評価は、日本語を母語とする同時通訳者によって、自身のこれまでの経験の観点を反映しつつ、Multidimensional Quality Metrics（MQM）[22, 23] を用いた評価を依頼した。その中から第 1 の実験結果に基づいて、30 語以上で、3 つ以上のアライメントが取れたトークンを持つ 172 のセグメント中 25 セグメントを選択し比較した。

N_{align}	字幕 (翻訳)	同時通訳
2	0.7087 (4614)	0.7297 (2269)
3	0.6933 (2823)	0.7242 (1014)
4	0.7188 (1598)	0.7772 (428)
5	0.7293 (945)	0.7941 (175)
6	0.7233 (510)	0.8054 (86)

表 2　提案手法による閾値 θ=0.71 の場合の SI と字幕の同期の程度。丸括弧内の数字は評価対象となったセグメント数を示す。

5.2　同時通訳と翻訳の比較分析

第 1 の実験にて、図 1 は異なるアライメントの閾値 θ での結果を示し、信頼できるアライメントの数と閾値のトレードオフから、表 2 にて θ=0.71 の場合の詳細を示している。具体的にはアライメントが取れた要素の数（N_{align}）を変化させ、データセット全体で異なる部分集合のスコアを比較している。その理由は 2 つある。1 つ目は、アライメントが取れた要素の数が順位相関計算の安定性に影響を与えるからである。2 つ目は、大きな N_{align} を持つつまり原発話が長いものは、短いものよりも語順の並び替えに影響を受けるからである。ここで注意しておくべきは、同じ N_{align} を持つスコアは、評価が可能な部分集合の違いにより直接比較できないことである。

結果から、N_{align} が大きいところでは、SI にて強い語順同期性が示された。これは、信頼性の高い単語のアライメントを多く含むセグメントつまり発話が長い SI において、原発話に対する目的発話の語順同期性が翻訳よりも高いことを示唆している。一方で、N_{align} が小さいところでは、翻訳と SI の間の語順同期性の差は小さかった。

以下に、これらの結果を例を挙げて具体的に議論する。表 3 は短いセグメントの例を示す。ここでは SI は翻訳と同程度の語順の並び替えが起きている。これは、同時通訳者が短い発話に対しては自然な日本語の語順で訳出することが可能なため、SI と

原発話書き起こし	(1) I learned / (2) new characters / (3) every day / (4) during the course of the next 15 years.
同時通訳書き起こし	(4) それから 15 年 (*during the course of the next* 15 *years*) / (3) 毎年ずっと (*every day*) / (2) 新しい文字を (*new characters*) / (1) 学んできました。(*I learned*)
字幕（翻訳）	(4) その後 15 年間 (*during the course of the next* 15 *years*) / (3) 毎日 (*every day*) / (2) 新しい漢字を (*new characters*) / (1) 習いました。(*I learned*)

表 3　原発話が短い例

原発話書き起こし	(1) Now mathematicians / (2) have been hiding and writing / (3) messages in / (4) the genetic code / (5) for a long time / (6) but / (7) it's clear / (8) they were mathematicians and not biologists / (9) because if you write long messages with / (10) the code / (11) that the mathematicians developed / (12) it would more than likely lead to / (13) new proteins being synthesized / (14) with unknown functions.
同時通訳書き起こし	(1) 数学者は (*Now mathematicians*) / (3) この様なメッセージを (*messages in*) / (4) 遺伝子コードで (*the genetic code*) / (2) 作って来たんです。 (*have been hiding and writing*) / (6) けどもしかし (*but*) / (11) 数学者は生物学者ではありません。(*Mathematicians are not biologists.*) そして間違ってる物もある訳です / (13) 新しいタンパク質を合成してしまう訳です。(*It means synthesizing a new protein.*)
字幕（翻訳）	(5) 長い間 (*for a long time*) / (4) 遺伝子コードに (*the genetic code*) / (3) メッセージを書き込む仕事は (*messages in*) / (1) 数学者が (*Now mathematicians*) / (2) 行ってきました。 (*have been writing*) / (8) 数学者は生物学者ではありません (*Mathematicians are not biologists*) / (11) 数学者が作成した (*the mathematicians developed*) / (10) コードを使って (*the code*) / (9) 長いメッセージを書いたとすると (*if you write long messages with*) / (14) 未知の機能を持った (*with unknown functions*) / (13) 新しいタンパク質の合成に (*new proteins being synthesized*) / (12) つながることでしょう。(*It would more than likely lead to*)

表 4　原発話が長い例

自動評価	Pearson の相関係数
COMET-QE	0.161
BERTScore(F1)	0.1111
Proposed	**-0.497**

表 5　長いセグメントにおける MQM ベースの人手評価と自動評価の相関比較

翻訳の間の語順同期性の差が小さいことを示唆している。対照的に、表 4 のような長いセグメントでは、SI は翻訳よりも原発話の語順同期性が高かった。これは、SI の時間制約と同時通訳者の記憶容量の制約により、SI では単語の大幅な並べ替えが困難であることに起因している。これは SI において省略が頻繁に起こっていることからも示唆される。翻訳は日本語的な流暢さを維持するため、より多くの単語の並べ替えが行われているものの、この翻訳の場合はセグメントの後半部分「because」で始まる部分から最初に戻すことがないため、一見原発話と同期しているように見える。これは、TED Talks での字幕を使用した翻訳スタイルに由来することから、これは一般的な翻訳とわずかに異なる。一般的な翻訳との違いは、表 2 に現れているようなスコアの差に結びつくはずである。最後に、語順同期性と通訳品質の人手評価との関係について調査を行った。表 5 は、自動評価指標のスコア（COMET-QE、BERTScore（F1）、および提案手法）と MQM エラースコアとの Pearson 相関係数を示している。提案手法は人手評価と負の相関を示した一方で、他の評価指標はほとんど相関が見られなかった。これは従来の機械翻訳の評価指標を用いて SI の品質を測定することの難しさを暗示しており、SI 評価において語順同期性を計る有用性を示している。SI・SiMT の評価では、従来の単語ごとの一対一対応の意味評価ではなく、リアルタイムの制約下で適切な内容伝達が要求される SI の特徴を踏まえたアプローチの必要性も示唆している。

6　おわりに

本研究では、順位相関係数に基づく語順同期性を BERTScore を用いて計る評価指標を提案した。提案手法は、英日 SI コーパスを用いた実験を通じて、長いセグメントにおいて、翻訳よりも SI の語順同期性が特に強いことを示した。今後の課題として、要約、言い換え、省略などを含む SI および SiMT により頑丈な語順同期性評価、および SI および SiMT に特化した品質評価などが挙げられる。

謝辞

本研究は JSPS 科研費 21H05054、22H03651 の助成を受けたものです。

参考文献

[1] He He, Jordan Boyd-Graber, and Hal Daumé III. Interpretese vs. translationese: The uniqueness of human strategies in simultaneous interpretation. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 971–976, San Diego, California, June 2016. Association for Computational Linguistics.

[2] Kosuke Doi, Katsuhito Sudoh, and Satoshi Nakamura. Large-scale English-Japanese simultaneous interpretation corpus: Construction and analyses with sentence-aligned data. In **Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)**, pp. 226–235, Bangkok, Thailand (online), August 2021. Association for Computational Linguistics.

[3] Zhongxi Cai, Koichiro Ryu, and Shigeki Matsubara. What affects the word order of target language in simultaneous interpretation. In **2020 International Conference on Asian Language Processing (IALP)**, pp. 135–140, 2020.

[4] Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor O.K. Li. Learning to translate in real-time with neural machine translation. In **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers**, pp. 1053–1062, Valencia, Spain, April 2017. Association for Computational Linguistics.

[5] Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3025–3036, Florence, Italy, July 2019. Association for Computational Linguistics.

[6] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In **Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing**, pp. 944–952, Cambridge, MA, October 2010. Association for Computational Linguistics.

[7] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In **International Conference on Learning Representations**, 2020.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 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. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[9] Jinming Zhao, Yuka Ko, Ryo Fukuda, Katsuhito Sudoh, Satoshi Nakamura, et al. NAIST-SIC-Aligned: Automatically-Aligned English-Japanese Simultaneous Interpretation Corpus. **arXiv preprint arXiv:2304.11766**, 2023.

[10] Kayo Matsushita, Masaru Yamada, and Hiroyuki Ishizuka. An Overview of the Japan National Press Club (JNPC) Interpreting Corpus. **Invitation to Interpreting and Translation Studies**, No. 22, pp. 87–94, 2020a.

[11] NAATI. Certified interpreter assessment rubrics, 2023. Accessed on 26.09.2023.

[12] European Union. Marking criteria for consecutive, 2023. Accessed on 26.09.2023.

[13] 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. Association for Computational Linguistics.

[14] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics.

[15] Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In **Proceedings of the Sixth Conference on Machine Translation**, pp. 1030–1040, Online, November 2021. Association for Computational Linguistics.

[16] Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. Monotonic infinite lookback attention for simultaneous machine translation. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 1313–1323, Florence, Italy, July 2019. Association for Computational Linguistics.

[17] Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. Overgeneration cannot be rewarded: Length-adaptive average lagging for simultaneous speech translation. In **Proceedings of the Third Workshop on Automatic Simultaneous Translation**, pp. 12–17, Online, July 2022. Association for Computational Linguistics.

[18] Elisa Robbe. **Ear-Voice Span in Simultaneous Conference Interpreting EN-ES and EN-NL: a Case Study**. PhD thesis, Ghent University, 2019.

[19] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. **Computational Linguistics**, Vol. 19, No. 2, pp. 263–311, 1993.

[20] Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 1627–1643, Online, November 2020. Association for Computational Linguistics.

[21] Zi-Yi Dou and Graham Neubig. Word alignment by fine-tuning embeddings on parallel corpora. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 2112–2128, Online, April 2021. Association for Computational Linguistics.

[22] Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. Multidimensional Quality Metrics (MQM): A Framework for Declaring and DescribingTranslation Quality Metrics. **Revista Tradumàtica: tecnologies de la traducció**, Vol. 12, pp. 455–463, 2014.

[23] Milind Agarwal, et al. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In **Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)**, pp. 1–61, Toronto, Canada (in-person and online), July 2023. Association for Computational Linguistics.