タグ付き混合データ学習と自己教師あり学習による同時通訳データを用いた End-to-End 同時音声翻訳

胡 尤佳　福田 りょう　西川 勇太　加納 保昌　須藤 克仁　中村 哲
奈良先端科学技術大学院大学
ko.yuka.kp2@is.naist.jp

概要

同時音声翻訳は、話し手の発話終了を待たずに漸進的に翻訳を行う技術である。高品質かつ低遅延の同時音声翻訳を実現するには、適宜省略や、話し手が話した内容をできるだけ早く訳出する等の、実際の通訳者の技法を同時通訳コーパスから学習することが有効であると期待される。本研究では、同時通訳データが不足している問題を軽減するため、同時通訳データに加えて話し手の発話終了を待ってから翻訳を開始する翻訳を前提としたオフラインデータと混合し、スタイルタグを用いて出力スタイルを区別する学習法を提案する。実験結果から、文の意味的類似度を測る評価における同時通訳テストデータでの翻訳性能向上の効果が、提案手法において示された。

1　はじめに

同時音声翻訳 (SimulST) は、文末を待たずに話し手の発話を漸進的に翻訳する技術であり、話し手の発話終了を待ってから翻訳を開始するオフライン音声翻訳 (Offline ST) と比較し、低遅延での翻訳を実現できる。近年ではニューラルネットワークに基づく End-to-end の SimulST モデルが提案されているが [1, 2]、大規模な同時通訳 (SI) データがなかったため、MuST-C [3] のような TED の字幕データからなるオフライン (Offline) データを用いて学習されてきた。SimulST は話し手の発話終了を待たずに漸進的に翻訳するという点で、通訳者が行う同時通訳の考え方に類似している。通訳者は、話し手の発話に追従するため、不必要な語の省略や、話し手が話した内容をできるだけ早く話すなどの技法を取り入れている。そのため、高品質で低遅延な同時音声翻訳を実現するには、SI データを使用して同時通訳の仕方を学習させることが有効であると期待される。

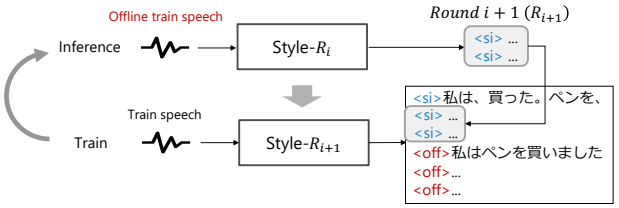

図 1　自己教師あり学習による多段階学習とタグ付き混合データ学習を組み合わせた手法

特に、英語と日本語はそれぞれ SVO 言語、SOV 言語であることから、Offline 翻訳においては訳出される語順の違いが大きい。Offline データを用いて同時翻訳を学習してしまうと、訳出時に原言語で始めの方に出てきた内容が目的言語だと後の方に出てくるなどの場合があり、遅延が大きくなる恐れがある。英日の言語対での SI データはいくつかあるものの [4, 5, 6, 7]、Offline データと比較して極めて少量である。本研究では、より多量に利用可能な Offline データも SI データの学習と同時に利用することで、SI データが少量である問題に取り組む。与えられた英語の原文 (Source) に対する Offline と SI の日本語の例を図 2 に示す。Offline と SI のスタイルには大きな違いがあり、図 2 では以下の違いがある。

- Offline ではほとんどの原文の語が日本語に翻訳されているが、SI ではいくつかの原文の語が省略されている。
- SI は Offline と比較し、文法的な流暢さが犠牲となっているが、原文の前半を後半より先に訳すことで、低遅延の翻訳を実現できる。

このような違いは、話し手の発話に追従できるように適宜省略を行いつつ、SI が通訳の同時性を重視していることから生まれる。したがって、SI データを用いてこうした翻訳を SimulST が学習することで、同時性の向上が期待できる。

学習方法として、SI データを用いてモデルを最初から学習する方法がまず考えられるが、データが不

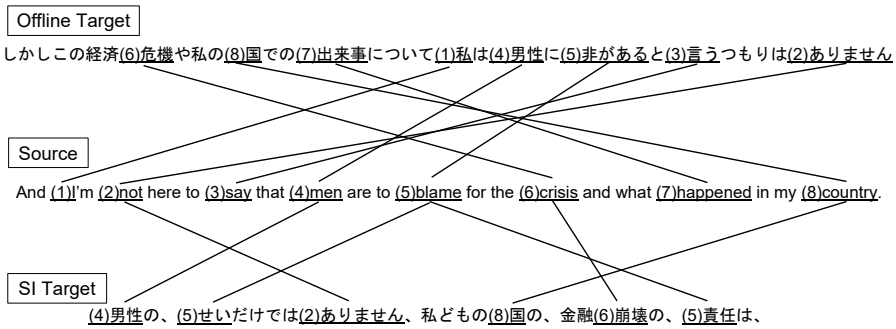

図 2　英日での Offline と SI の違いの例。図中の実線は単語の対応を表す。例文は NAIST-SIC-Aligned [4] に含まれる。

足しており、モデルの学習が上手くいかない恐れがある。また、多量の Offline データで学習された事前学習済みモデルに対して SI データを Fine-tuning させる方法が考えられる。しかし、多量の Offline データで学習された事前学習済みモデルに対してそのまま少量の SI データを Fine-tuning させると、過学習を起こしやすく、省略でない部分も含めて過剰に省略してしまうモデルが作成される恐れがある。

本研究では、SI または Offline 出力を明示的に意味するスタイルタグが付与された混合データを用いて単一の SimulST モデルを学習する手法を提案する。スタイルタグを付与することにより、SI または Offline スタイルの出力を選択的に生成するようデコード時にモデルに指示することができる。また本研究では、一度作成されたモデルから生成された擬似 SI データをさらに次の学習に用いる自己教師あり学習を多段階に行う手法も組み合わせ、SI データが少量である問題のさらなる軽減を目指した。実験の結果、提案手法により、文の意味的類似度を測る評価手法である BLEURT、COMET、COMET-QE での改善を示したことが分かった。また、出力結果を見た結果、提案モデルは従来モデルと比較し、より SI スタイルの出力が期待できることが分かった。

2　関連研究

本手法は、リソースの豊富な Offline 翻訳からリソースの乏しい SI へのドメイン適応と見なすことができる。ドメイン外モデルに対してドメイン内データを Fine-tuning させる手法は、いくつか考えられてきた [8, 9]。Chu らは、多量のデータに基づくドメイン外モデルを用いて直接少量のドメイン内データを学習させると、小さいドメイン内データに過剰適応 (Overfitting) する課題を、ドメイン内、外データであることを表すタグを付与することにより軽減した [9]。Tagged Back-translation [10] では、タグベースの手法を逆翻訳に基づく Data Augmentation に応用しており、NMT モデルの学習における逆翻訳ノイズに対処するため、オリジナルの原文と逆翻訳から得られた原文を区別するタグを付与している。本研究は、これらのタグベースの手法から着想を得ており、SI データの不足の課題に取り組む。

3　提案手法

3.1　タグ付き混合データ学習

本研究では、比較的多量な Offline データを用いて SI データの不足を補い、SimulST モデルを学習する方法を提案する。先行研究に従い [9]、学習では目的言語文字列の先頭にスタイルタグを付与し、推論時には最初に指定されたタグを Forced decoding し、タグが指すスタイルの出力の訳出を行う。本研究では、SI の場合は<si>、Offline の場合は<off>、の 2 種類のタグを用いる。

3.2　自己教師あり学習による多段階学習とタグ付き混合データ学習

SI データの不足をさらに軽減するために、本研究では擬似生成された SI データを用いて段階的に学習させる自己教師あり学習を、タグ付き混合データ学習の手法と組み合わせる。自己教師あり学習を利用したタグ付き混合データ学習の概要を図 1 に示す。まず、すでに学習された Style-R_i モデルに対して Offline 学習データの原言語音声を入力し、出力時に<si>を Forced decoding することで SI スタイルの擬似 SI データを生成させる。出力された擬似 SI データを、Round $i+1$ (R_{i+1}) データとする。次に、生成された R_{i+1} データに<si>タグを付与し、オリジナルの Offline、SI データに追加して、元モデルの Style-R_i モデルに対してタグ付き混合データ学習を

表 1　実験で利用した Offline と SI データのサイズ

	Train	Dev	Test
Offline	328,639	1,369	2,841
SI	65,083	165	511

行い、学習されたモデルを Style-R_{i+i} とする。この操作を多段階に繰り返すことで、擬似生成されるデータの質が向上していき、SimulST モデルの性能向上ができることが期待される。

4　実験

4.1　データセット

本研究で用いる Offline, SI データは両方とも、原言語音声、原言語テキスト、目的言語テキストにより構成される ST データを用いた。Offline データは MuST-C [3] v2 英日データ、SI データは NAIST-SIC-Aligned [4] [1]を用いた。NAIST-SIC-Aligned は、NAIST-SIC [7] を用いて、原言語テキストと目的言語テキストの文をアライメント、Filtering して構築された。NAIST-SIC-Aligned における、INTRA、AUTO-DEV、AUTO-TEST データをそれぞれ SI の学習、開発、評価データとした。これらの SI データのうち、英語の Forced aligner である Gentle [2]を用いて、SI データの英語のテキストと MuST-C の対応する音声とのアライメントを取れるものをデータとして利用した [3]。実験に使用したデータセットのサイズを表 1 に示す。最終的な評価には Offline データではなく、SI 評価データを用いた。

4.2　同時音声翻訳

Fine-tuning 前のモデルは Fukuda ら [11] の base モデルを用いている。base は、事前学習済みモデルである HuBERT を Encoder、mBART を Decoder として初期化されたモデルを複数の Offline ST データで Fine-tuning して作成されている。Fine-tuning については、言及がない限り [12] での Fine-tuning 時の設定に従う。文単位モデルと、[12] のように Prefix Alignment [13] をベースにしたモデルで比較した際、前者の方がベースライン手法の性能が高い傾向があったため、本研究では前者を採用した。Fine-tuning の際に以下の設定で実験を行ったものをそれぞれベースライン、提案手法とした。

ベースライン: Offline, SI　それぞれ、Offline, SI データで Fine-tuning したモデル。

提案手法 1: Style, Style-Up　タグ付き混合データで Fine-tuning したモデル (-Up は SI データを upsampling したモデル)。

提案手法 2: Style-R_0, Style-R_1, Style-R_2, Style-R_3　自己教師あり学習をタグ付き混合データ学習に適用したモデル。Style-R_3 は、R_0 から R_3 まで使って多段階学習させたモデル [4]。

また、Fukuda らと同様に、同時音声翻訳のための段階的な出力のためのデコード手法は Local Agreement (LA) [14] を採用した。それぞれのモデルにおいて、部分出力を得るための部分入力音声のセグメントサイズの単位を {200, 400, 600, 800, 1000}ms の 5 パターンでデコードを行い、遅延域のコントロールを行った。遅延評価では部分訳出の終了タイミングを考慮した遅延評価指標である Average Token Delay (ATD) [15] を用い、性能評価には、参照訳との単語一致度を測る BLEU、参照訳との出力文の文の意味的類似度を測る BLEURT [16]、参照訳、出力文だけでなく原言語参照訳との類似度も考慮した COMET [17], COMET-QE を用いた。

5　実験結果と考察

ベースラインと提案手法の遅延評価 (ATD) と性能評価 (BLEU, BLEURT, COMET-QE) の関係を図 3 に示す [5]。実験結果から、BLEURT, COMET-QE において、いずれの場合でも提案手法で性能向上がみられた。提案法において、少量の SI データを upsampling し、かさ増した場合の効果がこれらの評価手法では見られず、むしろ低下を招く場合があった。その一方で、自己教師あり学習による擬似 SI データを用いた場合には翻訳性能の向上が見られ、少量の SI データの upsampling と比較し、より多様性のある擬似 SI データを用いることの有効性が示された。自己教師あり学習による多段階学習を行った場合とそうでない場合を比べると、前者の方がより性能向上の傾向が見られた。

一方で、BLEU における評価では、提案手法ベースラインである SI が最も良い結果となり、BLEURT, COMET-QE の場合と傾向が大きく異なる。SI 評価

1）https://github.com/mingzi151/AHC-SI
2）https://github.com/lowerquality/gentle
3）本研究で用いたデータを以下で公開している。
https://dsc-nlp.naist.jp/data/NAIST-SIC/Aligned-ST
4）本研究では、R_0 データを生成する元モデルとして Style を用い、R_0 データを base に対して適用したモデルを Style-R_0 とした。
5）COMET での実験結果は、COMET-QE と傾向が類似しており、付録に添付した。

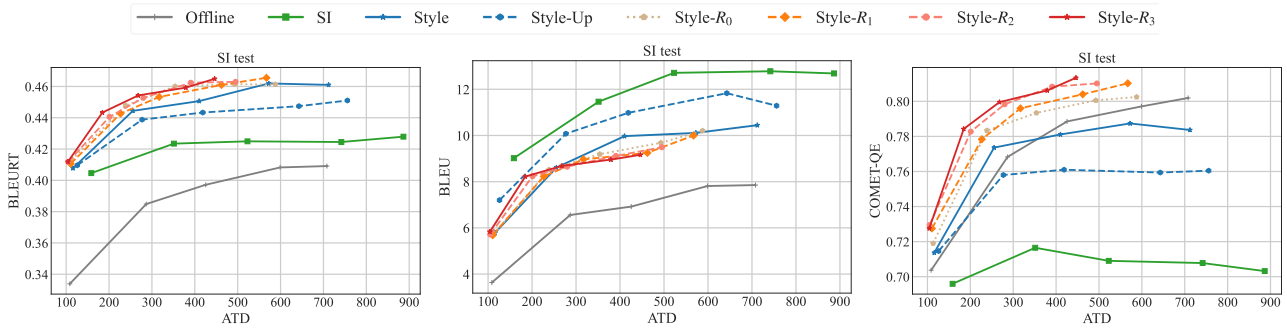

図 3　ベースラインと提案法における遅延評価 (ATD) と性能評価 (BLEU, BLEURT, COMET-QE) の関係

表 2　ベースラインと提案法における出力文と参照文の例 (入力音声のセグメント単位：400ms)

例文 1	
Source	It's probably the smallest of the 21 apps that the fellows wrote last year.
SI (ベースライン)	一番小さいアプリです。
Style-R_3 (提案手法)	これは、おそらく、21 のアプリの中で、最も小さいものです。昨年、フェローが書いたものです。
SI reference	昨年作ってくれたもので。
例文 2	
Source	It was running into bankruptcy last fall, because they were hacked into.
SI (ベースライン)	破産したんです。この秋に破産したんです。
Style-R_3 (提案手法)	それは、昨年、破産につながったものです。なぜなら、彼らは、不正に侵入されたからです。
SI reference	破産をしたのは、去年の秋なんです。ハッキングをされたからです、

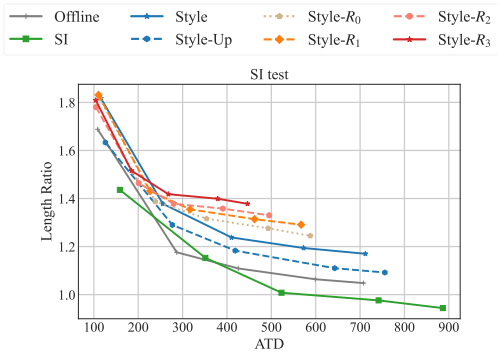

図 4　SI 評価データにおける参照訳と出力文の長さ比

データにおける出力文と参照訳との長さ比較を図 4 に示す。提案手法と比較すると、SI model からの出力長は短い傾向にあり、参照訳に最も近い長さの出力がされていることで、BLEU の向上に寄与したと考えられる。しかし、一部の SI 評価データに含まれる参照訳は、本来訳出すべき原言語音声の内容を過剰に削った内容で訳している場合があった。例文を表 2 に示す。例文 1 においては、SI model が過剰に省略を行っているものの、参照訳が過剰に短い。このような過剰に短い参照訳を含む SI 学習データで直接モデルが学習されたことで、SI model は過剰に省略を行うモデルになったと考えられ、例えば例文 2 の SI model 出力では、「they were hacked into」の部分が翻訳されず訳抜けになっている。

このような過剰に短い参照訳が、SI 評価データに多く含まれていたことから、参照訳のみに頼った評価が信頼できない恐れがある。そこで本研究では、COMET-QE のような、原言語テキストをベースにした評価も行った。COMET-QE での結果は、今回採用した評価指標の中で、それぞれの手法間の差が最も大きく見られた。特に、自己教師あり学習を取り入れた提案手法 2 においては、Style-R_0 から Style-R_3 へ多段階学習を繰り返すことにより、低遅延かつ高品質の翻訳がモデルで実現できることが期待される。また、図 4 より、提案法がベースラインと比べ出力が長く、自己教師あり学習による多段階学習を繰り返すほどその傾向がみられることが分かる。しかし実際の出力では、表 2 に示すように、「これは」、「なぜなら」、「彼らは」などの、内容に大きく影響しないつなぎ言葉や主語などがベースラインと比べ多く出力されている傾向があり、これらの語を除いても意味が通じるケースが多く、性能に大きく影響しないと考えられる。

6　おわりに

本研究では，同時音声翻訳において少量の SI データを用いた学習手法としてタグ付き混合学習を提案し、自己教師あり学習による多段階学習も取り入れた。実験結果から、より同時通訳らしい同時音声翻訳が提案手法により期待できることが分かった。今後の課題として、本研究でのモデルをベースに、通訳者が行うような適宜省略が可能なモデルの作成や、出力結果の人手評価を考えている。

謝辞

本研究の一部は JSPS 科研費 JP21H05054 と JST SPRING プログラム JPMJSP2140 の助成を受けたものである。

参考文献

[1] Xutai Ma, Juan Pino, and Philipp Koehn. SimulMT to SimulST: Adapting simultaneous text translation to end-to-end simultaneous speech translation. In **Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing**, pp. 582–587, Suzhou, China, December 2020. Association for Computational Linguistics.

[2] Yi Ren, Jinglin Liu, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. SimulSpeech: End-to-end simultaneous speech to text translation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 3787–3796, Online, July 2020. Association for Computational Linguistics.

[3] 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. Association for Computational Linguistics.

[4] 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.

[5] Hitomi Toyama, Shigeki Matsubara, Koichiro Ryu, Nobuo Kawaguchi, and Yasuyoshi Inagaki. CIAIR Simultaneous Interpretation Corpus. In **Proceedings of Oriental COCOSDA**, 2004.

[6] Hiroaki Shimizu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. Constructing a speech translation system using simultaneous interpretation data. In **Proceedings of IWSLT**, 2013.

[7] 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.

[8] Rico Sennrich, Barry Haddow, and Alexandra Birch. Controlling politeness in neural machine translation via side constraints. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 35–40, San Diego, California, June 2016. Association for Computational Linguistics.

[9] Chenhui Chu, Raj Dabre, and Sadao Kurohashi. An empirical comparison of domain adaptation methods for neural machine translation. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 385–391, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[10] Isaac Caswell, Ciprian Chelba, and David Grangier. Tagged back-translation. In **Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)**, pp. 53–63, Florence, Italy, August 2019. Association for Computational Linguistics.

[11] Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano, Yuka Ko, Tomoya Yanagita, Kosuke Doi, Mana Makinae, Katsuhito Sudoh, Sakriani Sakti, and Satoshi Nakamura. NAIST Simultaneous Speech Translation System for IWSLT 2023. In **Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT2023)**, pp. 330–340, 2023.

[12] Yuka Ko, Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano, Katsuhito Sudoh, and Satoshi Nakamura. Tagged end-to-end simultaneous speech translation training using simultaneous interpretation data. In Elizabeth Salesky, Marcello Federico, and Marine Carpuat, editors, **Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)**, pp. 363–375, Toronto, Canada (in-person and online), July 2023. Association for Computational Linguistics.

[13] Yasumasa Kano, Katsuhito Sudoh, and Satoshi Nakamura. Simultaneous neural machine translation with prefix alignment. In **Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)**, pp. 22–31, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics.

[14] Danni Liu, Gerasimos Spanakis, and Jan Niehues. Low-Latency Sequence-to-Sequence Speech Recognition and Translation by Partial Hypothesis Selection. In **Proc. Interspeech 2020**, pp. 3620–3624, 2020.

[15] Yasumasa Kano, Katsuhito Sudoh, and Satoshi Nakamura. Average Token Delay: A Latency Metric for Simultaneous Translation. In **Proc. INTERSPEECH 2023**, pp. 4469–4473, 2023.

[16] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. Bleurt: Learning robust metrics for text generation. In **Proceedings of ACL**, 2020.

[17] 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, 2020.

A　付録

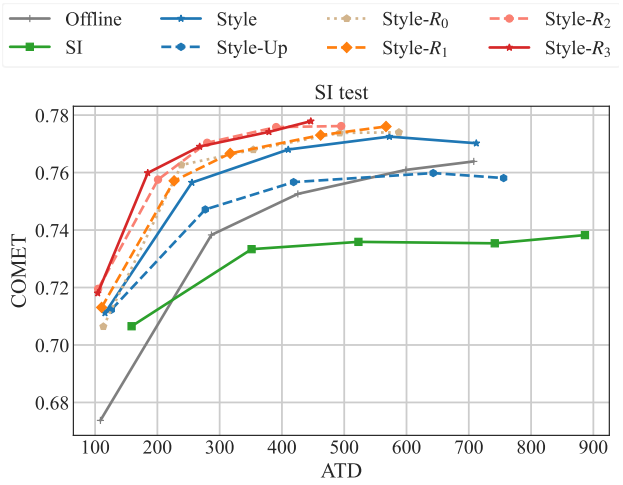

図 5　ベースラインと提案法における遅延評価 (ATD) と性能評価 (COMET) の関係