Simul-COMET: 原発話との語順差を考慮した同時通訳評価指標

土肥 康輔 [1]　蒔苗 茉那 [2]　坂井 優介 [2]　上垣外 英剛 [2]　渡辺 太郎 [2]
[1] 成蹊大学　[2] 奈良先端科学技術大学院大学
kosuke-doi@st.seikei.ac.jp,
{makinae.mana.mh2, sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp

概要

同時通訳者は原発話を短い単位に区切って処理することで，原言語と目的言語の語順の単調性を維持した訳出を行なっている．同時通訳文の自動評価には意味的類似性を考慮できる COMET 等の指標が用いられるが，COMET は語順の並び換えが多いオフライン翻訳データで学習されているため，同時通訳らしい単調性が高い文に低いスコアを付与するという課題がある．本研究では，単調性を考慮する品質評価指標 Simul-COMET を提案する．人手評価付きのオフライン翻訳データを，大規模言語モデルを用いて同時通訳調に変換し，Simul-COMET の学習に用いた．英日翻訳での実験の結果，Simul-COMET はオフライン翻訳よりも同時通訳調の翻訳に高いスコアを付与し，COMET よりもプロの同時通訳者による人手評価と高い相関を示した．

1　はじめに

同時通訳 (SI) は原発話をリアルタイムで翻訳する手法である．この時間的制約の中で高品質な翻訳を実現するため，同時通訳者は原発話を短い単位 (チャンク) に分割し，順次訳出を行なっている [1]．同時音声翻訳 (SiST) モデルも同様に，decoding policy によって現時点までの入力で翻訳を出力するか，追加の入力を待つかを決定している [2, 3]．その結果，SI 文は原発話文の語順に沿った，単調性が高い翻訳となる．図 1 の例では，オフライン翻訳の参照文 (off-ref) は原発話文 (src) と大きく異なる語順となっているのに対して，SI の参照文 (si-ref) は src の語順に沿った翻訳となっている．この単調性は，特に統語構造が大きく異なる言語対において，SI 文の品質評価の重要な側面である [4, 5]．

SiST モデルの翻訳品質は BLEU [6] や BLEURT [7], COMET [8] などの指標で評価されるが，参照文に off-ref が用いられることも多く，単調性は十分に評価されていない．参照文に si-ref を用いることで，BLEU や BLEURT はより正確に SI 文を評価できるようになるが，COMET は単調性の高い SI 調の翻訳 (si-tgt) よりもオフライン翻訳 (off-tgt) に高いスコアを付与するという課題がある [4]．

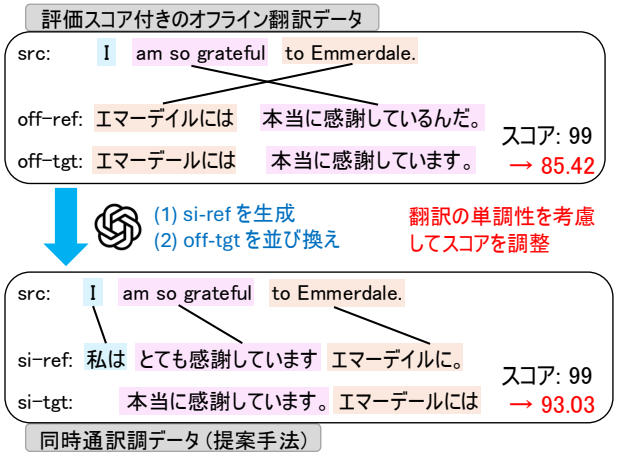

図 1　オフライン翻訳文と同時通訳文の例．オフライン翻訳データの参照文 (off-ref) と対象言語文 (off-tgt) は原言語文 (src) からチャンク順の入れ替わりがある．大規模言語モデルを用い，同時通訳調の参照文 (si-ref) は [5] の手法，対象言語文 (si-tgt) は off-tgt を並び替えることで作成する．

そこで本研究では，翻訳の単調性を考慮した品質評価指標 Simul-COMET を提案する．COMET は原言語文，人手評価付き対象言語文，参照文から学習されるニューラル評価指標であり，デフォルトモデル[1)] (wmt22-comet-da) は，オフライン機械翻訳モデルの出力を人間が 0〜100 点で評価した Direct Assessment (DA) データで学習されている．しかし，SI 調の翻訳に対する人手評価データの新規作成は高コストであるため，本研究では大規模言語モデル (LLM) を用いて，既存の DA データのオフライン翻訳文 (off-tgt) を内容と誤りを保持したまま SI 調の翻訳文 (si-tgt) に変換する (図 1)．また，DA スコアは対象言語文の原言語文に対する単調性の度合いに基

1)　https://huggingface.co/Unbabel/wmt22-comet-da

づき調整し，si-ref は [5] の手法により LLM を用いて作成する．

本研究では，統語構造の差が大きく，翻訳の単調性が他の言語対と比べて品質評価に大きく影響する英日間の SI [4, 5] を対象として Simul-COMET を構築する．実験の結果，Simul-COMET は off-tgt よりも si-tgt に対して高いスコアを付与し，wmt22-comet-da よりもプロ同時通訳者による人手評価と高い相関を示した．

2 関連研究

BLEU に代表される初期の翻訳品質の評価指標は n-gram の表層一致に基づくものであり，文字単位の一致を考慮する chrF++ [9] や大きな語順変化に対応する RIBES [10] などが提案された．しかし，これらの表層一致に基づく指標は人手評価との相関が低いことが指摘されており [11]，文脈埋め込みにより意味的類似度を捉える BLEURT や COMET 等のニューラル指標が提案されている．近年では，GEMBA [12] 等の LLM に基づく指標も登場しているが，WMT24 Metric Shared Task では，XCOMET [13] や MetricX [14] 等の事前学習済み言語モデルに基づく指標が依然として高い性能を示している [15]．

SiST モデルは遅延と翻訳品質の指標により評価されることが一般的だが，SiST に特化した指標が存在する遅延とは異なり，翻訳品質の評価には機械翻訳のために開発された上記の指標が用いられている．これらの指標は SiST の人手評価と相関するという報告 [16] がある一方で，原言語文を分割したり要約したりする SI 特有の方略や，語順の差異の影響を受けやすいという報告もある [4, 17]．また，原言語文と対象言語文の単語アライメントに基づき相関係数を算出することで，単調性を直接的に評価した研究 [5] がある．翻訳の単調性は，通訳・翻訳研究分野では同時通訳者の用いる方略との関係で盛んに研究されており [18]，SiST の評価指標への導入は重要な課題である．

3 Simul-COMET

本研究は翻訳の単調性を考慮した品質評価指標 Simul-COMET を提案する．SI 調の翻訳文から成る DA データ (SI-DA) を作成し，Simul-COMET モデルの学習に用いる．本研究では，大きな統語構造の差異のために SI が特に困難である，英日間の SI のための Simul-COMET モデルを構築する．

System

You are an experienced English-Japanese simultaneous interpreter. Your task is to generate a simultaneous interpreting sentence whose meaning is exactly equivalent to a given translation.

User

Instructions:
'Salami technique' in simultaneous interpreting refers to a technique where the interpreter breaks down the source language input into smaller, manageable segments that each contain enough information to be accurately interpreted. You will be given a pair of source and translation sentences.
1. Segment a given source sentence into chunks. Please make sure to segment it so that each chunk forms a meaningful unit for simultaneous interpreting.
2. Segment a given translation sentence based on Japanese "bunsetsu".
3. Align ALL the bunsetsus to source chunks without changing the order. Make sure that ALL the bunsetsus are aligned to a source chunk.
4. Save them in a list.

(Examples here)

Source:
Translation: Target data here

Output

```
{ "segmented_pair": [
    { "segmented_src": "It also",
      "segmented_tgt": [ "また、" ] },
    { "segmented_src": "did not take into account",
      "segmented_tgt": [ "考慮しなかった。" ] },
    { "segmented_src": "levels of indoor air pollution.",
      "segmented_tgt": ["室内空気汚染の", "レベルを" ] } ],
  "output_si": "また、/ 考慮しなかった。/ 室内空気汚染のレベルを" }
```

図 2 off-tgt を si-tgt に変換するプロンプトテンプレート

3.1 SI-DA データの構築

図 1 のように，wmt22-comet-da の学習データには src，off-ref，off-tgt の 3 種類の文が含まれる．はじめに，このオフライン DA データから si-ref と si-tgt を LLM を用いて作成する．次に，off-tgt と si-tgt が同一の意味となっていることを担保するためにフィルタリングを行う．

LLM には GPT-4o-mini[2] を [5] と同一設定のもと用い，オフライン DA データには WMT 2020 Metric Shared Task [19] から英日データ部分を使用する．

si-ref [5] の手法に基づき，LLM によって src をチャンクに分割し，それらをチャンク順に翻訳することで作成する．

2) https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

si-tgt　si-ref と異なり，off-tgt の内容と誤りを保持したまま SI 調に変換する必要がある．そのため，si-ref の作成では src のみを用いたが，si-tgt の作成には src と off-tgt を用いる．si-tgt を作成するプロンプトを図 2 に示す．はじめに，src と off-tgt をチャンクに分割する．本研究は英日間の SI が対象のため，src は [5] の手法，off-tgt は文節に基づいて分割する．次に，off-tgt の各チャンクを src のチャンクのひとつに対応づける．誤訳などのエラーのために明確な対応関係がない場合であってもいずれかの src のチャンクに対応づけるようにしている．最後に，OpenAI の Structured Output[3] の指定により JSON 形式で出力を得る．

フィルタリング　LLM は一般に高い指示追従能力を持つが，誤りが生じる場合もある [20]．前節の方法で作成した si-tgt が元の off-tgt と同一の意味を保持していることを保証するため，BERTScore [21] によるフィルタリングを行う．off-tgt を参照文とし，si-tgt との precision，recall，F1 スコアを算出し，いずれかが閾値 α を下回る場合，その文対を除外する．dev データ 500 文を人手で評価し，候補値{0.80, 0.85, 0.90, 0.95}を比較した結果，$\alpha = 0.90$ が最適となり，最終的に train / dev / test にそれぞれ 6,078 / 352 / 330 文を得た．

3.2　DA スコアの調整

src に対する単調性を DA スコアに反映するため，[0, 1] の範囲に正規化した DA スコア[4] $\boldsymbol{DA} = DA_1, \ldots, DA_i, \ldots, DA_I$ (I は文の数を表す) を以下の式に基づき調整した:

$$DA_i^m = DA_i - penalty_i^m \tag{1}$$

$$penalty_i^m = 0.25 \times (1 - MS_i) \tag{2}$$

DA_i^m は単調性スコア MS_i により調整された i 番目の文の翻訳品質スコアを表す．MS_i は [5] の手法に従い，src と対象言語文 (off-tgt または si-tgt) の単語アライメントに基づいて算出したスピアマンの順位相関係数を [0, 1] の範囲に変換した値である．src と対象言語文の語順が完全に一致する場合 MS_i は 1 となるため，語順差が大きいほど DA_i^m は元の DA_i より低くなる．DA Relative Ranks dataset [22] の作成ルールを参考に，ペナルティの最大値を 0.25 とした．

3)　https://platform.openai.com/docs/guides/structured-outputs

4)　COMET 学習のための DA スコアの準備は，公開されている手順を参照した: https://github.com/Unbabel/COMET/issues/131

3.3　モデルの学習

3.1，3.2 節で述べたデータを用い，Simul-COMET モデルを学習する．低資源言語や未対応言語へ既存 COMET を適応させる研究 [23, 24] を参考に，(1) モデルをスクラッチで学習する方法，(2) wmt22-comet-da をファインチューニングする方法の 2 種類の学習方略を以下の 3 種類のデータ条件下で行う:

(a) **Offline:** オフライン DA データの英日部分．対象言語文: off-tgt，参照文: off-ref.
(b) **SI:** 構築した SI-DA データ．対象言語文: si-tgt，参照文: si-ref.
(c) **Mixed:** 対象言語文のみ off-tgt と si-tgt を混ぜる．対象言語文: off-tgt + si-tgt，参照文: si-ref

これらの各データタイプに対して，2 種類の DA スコアを用いる: (i) オフライン DA スコア (da)，(ii) 単調性で調整した DA スコア (mono)．従って，2 種類の学習方略，3 種類のデータ条件，2 種類の DA スコア条件を組み合わせた合計 12 種類のモデルを学習する (表 1 参照)．

off-tgt と si-tgt に同一の DA スコア調整手法を用いているため，off-tgt のほうが si-tgt より低いスコアとなる．そのため，学習時に対象言語文のタイプを明示的にモデルに提示するタグは用いなかった．全ての設定において，モデルは平均二乗誤差を用いて 5 エポック学習し，dev セットにおけるケンドールの順位相関係数の値が最も高いチェックポイントを保存した．ハイパーパラメータは wmt20-comet-da の設定[5] に従った．

4　実験

4.1　Simul-COMET スコア

Simul-COMET が SI 調の単調性が高い翻訳文に，オフライン調の翻訳文よりも高いスコアを付与するかを検証するために，テストセットの off-tgt と si-tgt を用いて評価実験を行なった．いずれの場合でも，参照文には si-ref を用いた．

表 1 は，si-tgt に対するスコアから off-tgt に対するスコアを引いた差を示している．正の値はモデルが off-tgt より si-tgt に高いスコアを付与していることを表し，負の値はその逆を表す．デフォルトの COMET モデル (wmt22-comet-da)，およびオフライ

5)　https://github.com/Unbabel/COMET/files/8502021/wmt20-comet-da-hyper-parameters.zip

表 1 si-tgt と off-tgt に対する Simul-COMET のスコア差．全ての差は統計的に有意な差であった ($p < .05$)．

Models	Scratch	Fine-tuning
wmt22-comet-da	−0.0366	–
off-da	−0.0284	−0.0373
off-mono	−0.0328	−0.0286
si-da	−0.0083	−0.0109
si-mono	−0.0105	−0.0068
mix-da	−0.0073	−0.0083
mix-mono	**+0.0087**	**+0.0030**

表 2 人手評価の評価者間信頼性

Categories	Pre.	Rec.	F1	QWK
翻訳品質	0.7604	0.7984	0.7609	0.4488
流暢さ	0.9333	0.9300	0.9316	0.1540
単調性	0.8257	0.8313	0.8243	0.6260

ンデータで学習された off-da は，si-tgt よりも off-tgt に高いスコアを付与しており，これは [4] で報告されている結果と一致する．

しかし，mix-mono 設定を除く全ての Simul-COMET モデルも同様に，off-tgt に高いスコアを付与した．si-mono 設定の結果は，SI 調のデータのみで Simul-COMET を学習することは不十分であることを示唆している[6]．Simul-COMET モデルが off-tgt よりも si-tgt に高いスコアを付与するようになるには，学習時に両タイプの文を使用し，それらのスコア差を学習することが必要であると考えられる．

4.2 人手評価

Simul-COMET が SI の人手評価と整合することを示すために，プロの同時通訳者による実際の英日間の SI 文を用いた評価を行なった．日本記者クラブで行われたアウンサンスーチー氏の記者会見[7]から 243 文を抽出し，記者会見の SI を担当した同時通訳者とは別の，20 年以上の経験を有するプロの同時通訳者 2 名による人手評価を行なった．評価にあたっては，Multidimensional Quality Metrics (MQM)[8] に基づく評価ガイドラインを作成し，翻訳品質，流暢さ，単調性の 3 観点を独立して評価した．

3 つの評価観点において，大多数の文 (約 75%以上) は「エラーなし」という評価を受けており，本 SI 文が比較的高品質であったことを示唆している[9]．多クラス分類の precision, recall, F1 スコアと 2 次重み付きカッパ係数 (QWK) を用いて算出した評価者間信頼性を表 2 に示す．F1 スコアは比較的高く，全ての評価観点で 0.75 を超えているのに対して，QWK は翻訳品質と流暢さで低い値となっている．単調性は src との語順の比較で比較的客観的に評価できるのに対して，他の 2 観点は主に評価者の主観に基づく評価であったためと考えられる．

表 3 Simul-COMET スコアと人手評価の間のスピアマンの順位相関係数．Avg. score は Rater A と B のスコア平均との相関係数である．* は相関係数が統計的に有意であることを示す ($p < .05$)．

Model	Rater A	Rater B	Avg. score
wmt22-comet-da	0.1250	0.0874	0.1324*
mix-mono (scratch)	0.0817	0.0893	0.0903
mix-mono (fine-tuning)	**0.1418***	**0.1556***	**0.1719***

人手評価との相関係数の算出には，off-tgt より si-tgt に高いスコアを付与することができた mix-mono 設定の Simul-COMET モデルと，デフォルトモデルの wmt22-comet-da を用いた．スピアマンの順位相関係数の結果を表 3 に示す．mix-mono (fine-tuning) が最も高い相関を示し，wmt22-comet-da が続いた．mix-mono (scratch) の相関係数は統計的に有意でなかった．Paird bootstrap resampling [25] で mix-mono (fine-tuning) と wmt22-comet-da を比較したところ，10,000 回の試行中 9,195 回で mix-mono (fine-tuning) が上回り，90%信頼区間は [0.004, 0.160] であった．また，mix-mono (fine-tuning) は Rater A と B のスコアの両方と有意な相関があった．これらの結果は，Simul-COMET がデフォルトの COMET モデルに比べ，人手評価との相関が高いことを示唆している．

5 おわりに

本研究では，翻訳の単調性を考慮した品質評価指標 Simul-COMET を提案した．英日翻訳での実験の結果，off-tgt と si-tgt を混ぜて学習したモデルが，SI 調の文に高いスコアを付与することを示した．また，Simul-COMET はデフォルトの COMET モデルよりも，SI の人手評価と高い相関を示した．今後は，より多くの言語対で提案手法の有効性を確認することが課題である．また，本研究では SI の単調性のみに焦点を当てたが，要約や省略といった他の側面も考慮できるようにすることで，人手評価との相関係数の値を高めていくことも今後の課題である．

6） 各設定での学習の効果は付録 A を参照．
7） https://www.jnpc.or.jp/archive/conferences/34352/report#
8） https://themqm.org/
9） 詳細は付録 B を参照

謝辞

本研究の一部は JSPS 科研費 JP21H05054 の助成を受けたものである．

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 **Proc. of NAACL**, pp. 971–976, 2016.

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

[3] Sara Papi, Marco Turchi, and Matteo Negri. AlignAtt: Using Attention-based Audio-Translation Alignments as a Guide for Simultaneous Speech Translation. In **Proc. of INTERSPEECH**, pp. 3974–3978, 2023.

[4] Kosuke Doi, Yuka Ko, Mana Makinae, Katsuhito Sudoh, and Satoshi Nakamura. Word order in English-Japanese simultaneous interpretation: Analyses and evaluation using chunk-wise monotonic translation. In **Proc. of IWSLT**, pp. 254–264, 2024.

[5] Mana Makinae, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. Simul-MuST-C: Simultaneous multilingual speech translation corpus using large language model. In **Proc. of EMNLP**, pp. 22185–22205, 2024.

[6] 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.

[7] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In **Proc. of ACL**, pp. 7881–7892, 2020.

[8] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In **Proc. of EMNLP**, pp. 2685–2702, 2020.

[9] Maja Popović. chrF++: words helping character n-grams. In **Proc. of WMT**, pp. 612–618, 2017.

[10] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In **Proc. of EMNLP**, pp. 944–952, 2010.

[11] Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. A survey of evaluation metrics used for nlg systems. **ACM Computing Surveys**, Vol. 55, No. 2, pp. 1–39, 2022.

[12] Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In **Proc. of EAMT**, pp. 193–203, 2023.

[13] Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. xCOMET: Transparent machine translation evaluation through fine-grained error detection. **arXiv**, Vol. arXiv:2310.10482, , 2023.

[14] Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, **Proc. of WMT**, pp. 756–767, 2023.

[15] Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In **Proc. of WMT**, pp. 47–81, 2024.

[16] Dominik Macháček, Ondřej Bojar, and Raj Dabre. MT metrics correlate with human ratings of simultaneous speech translation. In **Proc. of IWSLT**, pp. 169–179, 2023.

[17] Shira Wein, Te I, Colin Cherry, Juraj Juraska, Dirk Padfield, and Wolfgang Macherey. Barriers to effective evaluation of simultaneous interpretation. In **Findings of EACL**, pp. 209–219, 2024.

[18] Zhongxi Cai, Koichiro Ryu, and Shigeki Matsubara. What affects the word order of target language in simultaneous interpretation. In **Proc. of IALP**, pp. 135–140, 2020.

[19] Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2020 conference on machine translation (WMT20). In **Proc. of WMT**, pp. 1–55, 2020.

[20] Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. Revisiting compositional generalization capability of large language models considering instruction following ability. In **Proc. of ACL**, pp. 31219–31238, 2025.

[21] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **Proc. of ICLR**, 2020.

[22] Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In **Proc. of WMT**, pp. 62–90, 2019.

[23] Júlia Falcão, Claudia Borg, Nora Aranberri, and Kurt Abela. COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque. In **Proc. of LREC-COLING**, pp. 3553–3565, 2024.

[24] Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In **Proc. of ACL**, pp. 14210–14228, 2023.

[25] Philipp Koehn. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, **Proc. of EMNLP**, pp. 388–395, 2004.

A　Simul-COMET の各設定での学習の効果

DA-SI データを用いた学習の効果を検証するため，si-tgt が off-tgt より高いスコアを得た割合 (win rate) を算出した．表 4 に示すように，si-*モデルは，デフォルトモデルや off-*モデルより高い win rate を示し，SI 調データのみを用いた学習が限定的ながら正の効果を持つことが示唆された．また，off-tgt と si-tgt を混合しただけの mix-da 設定でも，デフォルトモデル，off-*モデル，si-*モデルより win rate が高くなった．一方，最も高い win rate を示した mix-mono モデルの結果から，si-tgt に一貫して高いスコアを与えるには，off-tgt と si-tgt を混合することに加え，単調性に基づくスコア調整が不可欠であることが示唆される．

表 4　si-tgt が off-tgt よりも高いスコアを付与された例の割合

Models	Scratch	Fine-tuning
wmt22-comet-da	0.1429	–
off-da	0.1492	0.1134
off-mono	0.1371	0.1539
si-da	0.3388	0.1984
si-mono	0.2254	0.2520
mix-da	0.3443	0.3169
mix-mono	0.6748	0.6531

B　プロの同時通訳者による人手評価結果

作成した評価ガイドラインに基づき，各 SI 文は評価観点ごとに 4 段階で評価された: none: エラーなし; minor: 軽度のエラー; major: 重要のエラー; critical: 深刻なエラー．図 3 は人手評価の詳細な結果を示しており，すべての評価観点において，多くの文が none と判定されたことが分かる．流暢性と単調性については，2 名の評価者が各ラベルをほぼ同様の割合で使用しており，流暢性では約 95% が none，約 5%が minor，単調性では約 75%が none，約 23%が minor，約 2%が major または critical と評価された．しかし，正確性は評価者間で判定の厳しさに差が見られ，一方の評価者は約 78%のセグメントに none を付与したのに対し，もう一方の評価者は約 90%に none を付与しており，評価の厳しさに違いがあることが示唆される．

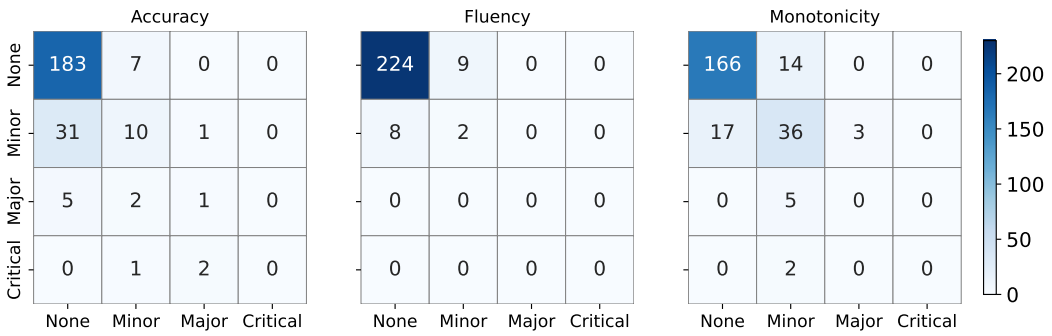

図 3　2 名のプロの同時通訳者による人手評価の詳細