固定区間チャンキングを用いた文書翻訳手法の提案と分析

王 小天 [1] 林 有源 [2] 申 展 [1] 谷中 瞳 [1,3,4]
[1] 東京大学 [2] 京都大学 [3] 理化学研究所 [4] 東北大学
{eternaleden0321, shenzhan, hyanaka}@g.ecc.u-tokyo.ac.jp
lin.youyuan.73v@st.kyoto-u.ac.jp

概要

近年の LLM の文書翻訳では, チャンクによって, 入力長への変動の適応が試みられているが, 訓練時と推論時の入力長分布の不一致が課題になる. これを解決するため, 本研究では固定区間チャンキング (FRC) による文書翻訳を提案する. FRC は動的計画法を用いて各チャンクの長さを規定の範囲内に収め, 訓練・推論の両段階で文書長を正規化する. 本稿では 5 つのデータ形式にわたる 3 つの訓練戦略を提案・分析する. 実験の結果, FRC ベースの手法は元モデルや既存翻訳手法を顕著に上回り, 文書翻訳ベンチマーク (IWSLT2017, BWB) において商用 LLM に匹敵する精度を達成した.

1 はじめに

現在, 長いコンテキストウィンドウを備えた大規模言語モデル (LLM) により, 文書翻訳の障壁であった入力長の制限は大幅に克服されつつある. しかし, 入力長の変動への適応, 翻訳テキストにおける流暢性などは, LLM ベースの文書翻訳タスクにおいても依然として克服すべき主要な課題である [1, 2, 3].

LLM ベースの研究初期には, 主にプロンプティング技術が探索された [4, 5]. Wu ら [5] は文単位で逐次入力を行い, 先行するソース文および翻訳結果をコンテキストとして利用する手法を提案した. 彼ら [5] は 7B モデルのフルファインチューニングなどを試みたが, その精度は当時の SOTA システムや GPT-4 等の商用モデルに及ばなかった. 一方, 文単位からチャンク単位へと移行し, 多様な入力長の分布への適応するため, Ramos ら [6] は, 各文書を $k \in \{1, 2, 4\}$ 固定数のチャンクに分割, 先行するソース文と翻訳済みテキストをコンテキストとして利用する「chunk-to-chunk」訓練パラダイムを提案したが, 訓練とテストの入力長の分布を厳密に一致させるには至っていない. その結果, 推論時に未学習の入力長に直面した際の精度低下が懸念されている [2]. また, Blob [7] のような既存の文積み上げ手法は, 固定長のスライディングウィンドウや貪欲法的な戦略に依存しているため, 長さ分布の一貫性を保証する大域的最適化が困難であり, 小さな断片的なチャンクを生成しやすいという欠点がある.

これらの問題を解決するため, 本研究では動的計画法 (DP) を適用した固定区間チャンキング (**FRC**: **F**ixed-**R**ange **C**hunking) を提案する. 本手法は, すべての文書チャンクの長さを所望の範囲内に収めるよう制御し, 大域的に最適されたセグメンテーションを実現する. これにより, 訓練と推論の入力長の分布を一貫して整合させることが可能となる. さらに, 文アライメントが不完全な文書レベル並列コーパスにも対応可能な, 二重境界マッチングに基づく軽量なチャンク対応付け手法を提案する. 最後に, 5 種類の訓練データを構築し, これらを 3 つの訓練戦略に適用して 4 種類のモデルを訓練する. その上で, これら各モデルの特性について, 複数の評価指標を用いて詳細な調査および比較検証を行う.

2 訓練データの構築

2.1 固定区間チャンキング

文書から予め分割されたユニットの集合 $U = \bigcup_{i=1}^{n} u_i$ を, 固定区間 $[m, M]$ のチャンクへと分割する処理を最適化する 2 段階動的計画法アルゴリズムを提案する. まず, コスト関数 $C(l)$ を設計した. この関数は, 長さを区間 $[m, M]$ 中央へ収束させつつ, 境界条件への違反量を最小化するように作用する.

$$C(l) = \begin{cases} (l - \frac{m+M}{2})^2, & \text{if } l \in [m, M] \\ min\{(l-m)^2, (l-M)^2\}, & \text{if } l \notin [m, M] \end{cases} \quad (1)$$

前処理 M トークンを超える単一ユニット u_i は DP プロセスの前にあらかじめ分離する.

段階 1: 制約最適化 まず, 全てのチャンク長 l が厳密に $m \le l \le M$ を満たす分割の探索を行う. 最初

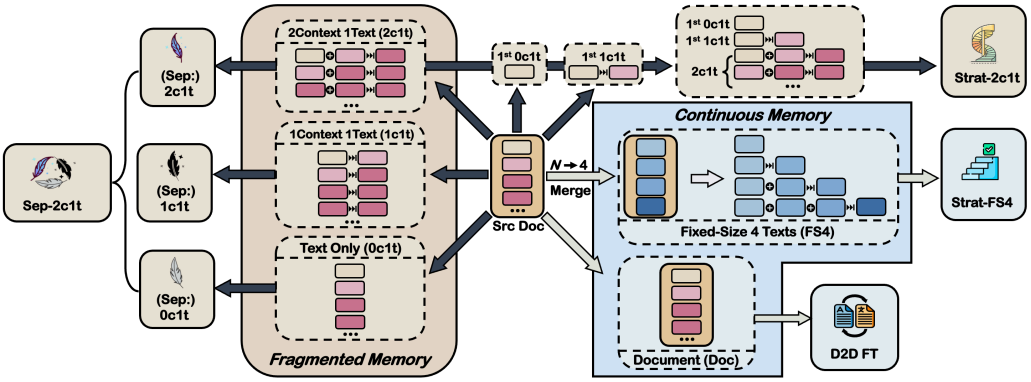


図 1: 同一の文書から派生した, 4 つのモデル構成に向けた訓練データの構築スキーム.

の i 個のユニットをパッキングする際の最小コストを $f[i]$ と定義する. $f[n] < \infty$ であれば, この分割を最適な厳密分割として採用する.

段階 2: 緩和最適化　厳密な分割が不可能な場合, 制約を緩和する. 本手法では, 最大 K 個のチャンクまで $[m, M]$ の境界違反を許容する. DP の状態を $f[v][i]$ へと拡張する. これは, 先頭から i 個のユニットを対象とし, 境界条件違反を正確に v 回用いた場合の最小コストを表す. 最終的な解は, $v \in \{1, ..., K\}$ に対する $f[v][n]$ の最小値として得られる.

2.2　チャンク対応付け

文書ペアにおいて完全な文アライメントの確保が困難であるという実用上の制約を想定し [8], 本研究では従来の文単位のアライメント手法に依存せず, チャンク単位の軽量なアライメント構築手法として二**重境界**マッチング戦略を提案する. その詳細は, 付録のアルゴリズム 1 に示す. まず, スコア関数 SimwRP を定義する. これは, LaBSE モデル [9] により得られる埋め込みに基づくコサイン類似度から算出される意味的類似度と, 相対位置情報を集約した指標であり, ハイパーパラメータ λ および σ によって両者の寄与が調整される. 各ソースチャンク c_i に対して, その終端境界 c_i^{end} に基づき, すべてのターゲット言語のユニット $t_j \in \mathcal{T}$ に対する初期マッチングスコアを算出する. 二重検証のため, 上位 k 個の候補を特定し, 後続チャンクの開始境界 c_{i+1}^{first} の対応するマッチング結果を組み込むことで, 各候補のスコアを再評価する. 続いて, DP を用い, 合計スコアを最大化する最適なパスを探索する. 境界制約が満たされない場合には, 実行可能なアライメントが得られるか, 或は探索空間が尽きる (すなわち $k = |\mathcal{T}|$ となる) まで, 候補ウィンドウサイズ k を段階的に拡大する. 最終的に構築された FRC ペアは, 後続の訓練データ生成プロセスの基盤として用いられる.

3　訓練方式

2 節で詳述した FRC ペアに基づき, 本研究では一連の訓練戦略を設計する. 図 1 に示すように, 以下の 5 種類の入力形式を定義する: (1) **Doc**: オリジナルの文書; (2) **0c1t**: 単一のチャンクを用いる形式; (3) **1c1t**: 直前の 1 チャンクをコンテキストとして付加する形式; (4) **2c1t**: 直前の 2 チャンクをコンテキストとして付加する形式; (5) **FS4**: チャンクを 4 つのセグメントに統合し, 先行するすべてのチャンクを逐次的にコンテキストとして利用する形式である. これらの形式を活用し, 3 つのタイプに分類される 4 つのモデルを訓練する: **D2DFT**: 文書ペアを用いて直接ファインチューニングを行うベースラインモデル; **Sep**: 0c1t, 1c1t, 2c1t の各形式を用いて個別に訓練された独立したモデル群. これらの組み合わせによる実装を Sep-2c1t と呼ぶ; **Stair**: 逐次的に階層化されたデータを利用する Stair-FS4 と, 0c1t および 1c1t のサンプルを用いて訓練を開始し, その後 2c1t サンプルへと移行する Stair-2c1t が含まれる.

4　実験設定

データセット　5 つの異なるコーパスから構成される DocBlocks データセット [6] を訓練データとして利用する. 具体的には, IWSLT [10, 11], BWB [12], GuoFeng [13], News Commentary, および Europarl [14] から構成される. 評価には, IWSLT2017 [11] および BWB のテストデータを用いる. 後者は, 6 つの Web 小説から抽出された連続する章によって構成されている. 各章の長さが比較的短いことを考慮し, 同一小説内の隣接する 2 章を連結して使用する.

表 1: 様々なデコーディング形式における各モデルの翻訳結果を示す. モデルの訓練設定と一致するデコーディング形式は, **太字**で強調している. 3 つの評価指標において, 最高性能の値は**太字**で示す. モデルごとに異なるデコーディング形式間で比較を行い, 最も高い値を下線で示す.

Models	Decoding Format	IWSLT2017 en-xx			IWSLT2017 xx-en			BWB zh-en		
		d-BLEU	d-COM.	GEM.-DA	d-BLEU	d-COM.	GEM.-DA	d-BLEU	d-COM.	GEM.-DA
Large-Scale LLMs										
GPT-4.1	d2d	36.53	86.23	94.72	39.19	85.79	91.93	22.85	81.06	94.26
Gemini-2.5-Pro	d2d	40.05	86.38	94.79	49.67	86.22	94.38	21.53	80.72	94.74
Qwen2.5-7B-Instruct										
Orig Model	d2d	25.88	80.00	61.16	35.54	84.54	84.46	18.57	80.20	81.21
	0c1t	30.00	81.83	63.73	35.42	85.03	84.34	19.38	80.72	79.00
	1c1t	29.80	82.05	63.72	35.29	84.94	84.64	19.07	80.47	79.00
	2c1t	30.07	82.42	64.33	35.37	84.88	84.60	18.75	80.59	79.03
	FS4	28.29	81.06	63.14	35.56	83.33	84.57	18.81	80.27	79.56
w/ D2DFT	**d2d**	29.17	82.38	74.31	22.46	82.78	80.07	26.41	80.63	84.03
	0c1t	32.07	82.76	74.03	35.78	83.99	76.24	25.09	79.92	82.36
STAIR-FS4	d2d	31.42	83.52	78.57	39.20	85.36	84.56	26.07	80.80	85.46
	0c1t	30.43	78.51	66.29	42.00	85.06	82.96	26.20	80.71	83.95
	1c1t	32.70	80.82	71.79	43.58	85.65	84.65	26.40	80.71	83.46
	2c1t	33.05	81.16	73.02	43.75	**85.73**	84.82	26.42	80.70	84.13
	FS4	30.68	81.57	75.62	37.30	83.94	84.49	26.40	80.49	84.18
STAIR-1c1t	0c1t	32.44	80.61	71.67	44.40	85.25	81.98	**27.14**	80.94	85.23
	1c1t	36.72	83.25	79.03	46.87	85.62	84.06	**27.14**	**81.10**	**86.38**
STAIR-2c1t	0c1t	32.06	79.79	69.74	43.51	85.27	80.95	26.84	80.76	84.33
	1c1t	36.74	83.00	78.38	45.65	85.62	84.01	26.77	80.89	84.97
	2c1t	36.97	83.08	78.01	46.24	85.70	84.19	26.71	80.88	85.28
SEP-2c1t	**0c1t**	38.21	84.13	81.67	45.21	84.83	82.89	26.28	80.47	84.15
	1c1t	**38.51**	**84.70**	**82.42**	**47.07**	85.57	**84.90**	26.58	80.72	84.74
	2c1t	38.32	84.59	82.40	46.80	85.47	84.86	26.44	80.64	84.97

モデルと訓練データ構築 バックボーンモデルとして QWEN2.5-7B-INSTRUCT [15] を採用し, ユニット長の測定基準として同モデルのトークナイザを利用する. 文書のユニット分割には Stanza [16] を用い, FRC のサイズは [256, 512] に設定する.

評価指標 d-BLEU [17, 18], SLIDE 方式 [19] に基づく d-COMET [20], LTCR (Lexical Translation Consistency Ratio) [21], および GEMBA-DA [22] を用いる. このうち, d-COMET については変種を提案しており, その詳細は付録 B に示す.

LTCR [21] は, 翻訳における用語の一貫性を評価する指標である. 本研究では, 翻訳候補と参照訳の間で LTCR を算出する改良案を提案する. 参照訳文内から spaCy を用いて特定された用語のみに基づいてスコアを算出する.

5 結果

QWEN2.5-7B-INSTRUCT モデルをベースとして訓練された各モデルの結果を表 1 に示す. 現在広く利用されている商用 LLM である GPT-4.1 [23] および Gemini-2.5-Pro [24] の評価結果も併記する.

全体像 オリジナルモデルにおいて, 多くの設定においてチャンキング手法が doc2doc (d2d) デコーディング戦略を概ね上回る性能を示した. さらに, 4 種類のファインチューニング済みモデルはいずれも, 分布内テストデータにおいてオリジナルモデルを一貫して凌駕している. IWSLT2017 en–xx 翻訳タスクでは, 最良モデルは 0c1t シナリオでオリジナルモデルから約 8 d-BLEU ポイントの改善を達成した. この傾向は他のタスクでも確認され, BWB zh–en タスクでは 7.76 ポイント, IWSLT xx–en タスクではさらに大きく 11.65 ポイントの改善が観測された.

モデルごとの分析 **D2DFT**: 0c1t は d2d デコーディング戦略と比較して IWSLT2017 においてより高い d-BLEU スコアを示した. d-BLEU は長文の評価に比重をおく指標であることから, 長文翻訳においてチャンキングが効果的であることを示唆している.

STAIR-FS4: 文書を厳密に 4 等分するのではなく, FRC 適用後にチャンクのマージする戦略を採用している. これによりチャンク長分布が離散化され, 訓練時と推論時における系列長が一致する確率が高まった可能性がある. 指標の総合評価では IWSLT2017 で

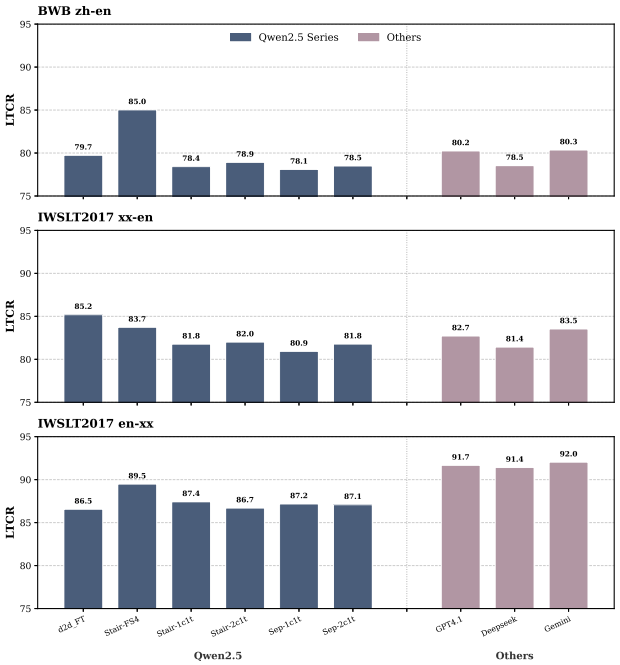
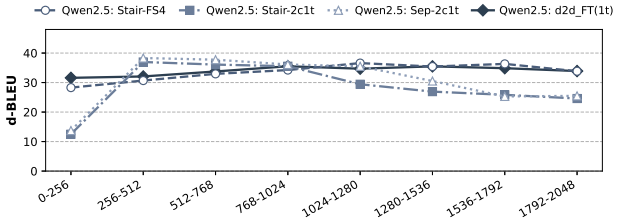
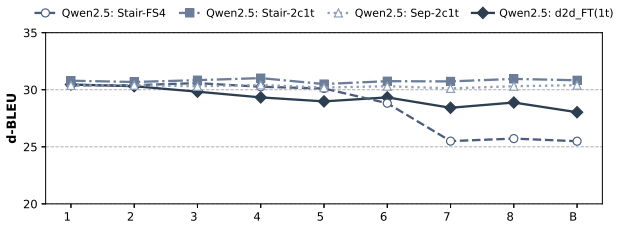


図 2: 各モデルの LTCR スコア.



(a) FRC 長さ区間の増大に伴う IWSLT2017 en–xx の d-BLEU スコア.



(b) 隣接する章の段階的な連結に伴う BWB zh–en の d-BLEU スコア.

図 3: 長さの変動に関する推論時分析.

d2dFT を上回り, BWB で同等の性能を示した.

Stair-1c1t / 2c1t: 訓練データにおいて最も多く含まれる入力形式をデコーディングに用いた際, 最良の性能を得られた. d-BLEU では両モデルはいずれも Stair-FS4 を一貫して上回る. 1c1t と 2c1t の比較では, 多くの場合 Stair-1c1t が僅差で優位であった.

Sep-2c1t: 独立して訓練された 3 つのサブモデルにより, 固定区間チャンキングを行う全ての設定で高い性能を発揮し, IWSLT2017 では全モデル中で最良の結果を達成している. 一方で, BWB テストデータにおける性能は Stair モデルを下回った. 総じて, Stair と Sep は d-BLEU/d-COMET で, 分布内テストデータにおいて, 商用 LLM に匹敵するが, GEMBA-DA では依然として大きな差が確認される.

LTCR の分析　図 2 に示すように, 各モデルの LTCR スコアを算出した. その結果, 完全な先行コンテキストを参照可能な d2dFT および Stair-FS4 モデルは, 高い LTCR スコアを示した. 特に後者は, 商用 LLM と同程度, あるいはそれを上回る性能を達成している. 一方, 他のモデルは自動評価指標において高いスコアが得られる場合があるものの, 参照可能なコンテキスト範囲が限定的であるため, 用語の一貫性は相対的に低下する傾向が観測された.

推論時チャンキング　図 3 に示すように, チャンク長に関する 2 つの観点から分析を行う. 第一に, 異なる FRC 長さ区間における推論時の性能を調査する. 第二に, BWB テストデータにおいて同一小説内の隣接する章を段階的に連結することで, テストデータ自体の長さを増加させ, 最終的に一冊レベルのデータを作成する. この手順により, 全 6 冊の小説について, 翻訳された全章を連結した状態で一冊レベルの評価を実施する.

Stair-FS4 と d2dFT モデルでは, FRC の増大に伴って d-BLEU が向上する傾向が確認された. しかし, テストデータの長さが一定の閾値を超えると, 性能は低下に転じる. この結果は, 両モデルが有効な処理長の範囲内では高い汎化性能を示すものの, その範囲が訓練データによって制約されていることを示唆している. 一方で, Stair-2c1t および Sep-2c1t モデルは, FRC が訓練時の設定 ([256, 512]) と一致している条件下で最高の精度を達成した. このことから, FRC を訓練時に用いた長さ区間に一致させることで, テストデータの長さによらず比較的頑健な性能が得られることが示唆される.

6　おわりに

本研究では, LLM ベースの文書レベル機械翻訳における入力長の変動と, 訓練・推論時の長さ分布の不一致に対処するため, 固定区間チャンキングを提案した. 動的計画法を用いることで, チャンク長を特定の範囲内に収める最適化を実現し, 二重境界マッチングによるチャンク対応付け手法を導入した. 実験の結果, Qwen2.5-7B-Instruct モデルにおいて, FRC ベースの手法は元モデルや d2dFT を顕著に上回り, 主要な商用 LLM に匹敵する精度を達成した.

謝辞

本研究は JST CREST, JPMJCR2565 の支援を受けたものである.

参考文献

[1] S. Liu, C. Lyu, M. Wu, L. Wang, W. Luo, K. Zhang, and Z. Shang. New trends for modern machine translation with large reasoning models, 2025.

[2] Z. Peng, R. Bawden, and F Yvon. Investigating length issues in document-level machine translation. In **Proc. 20th MTSummit**, pp. 4–23, 2025.

[3] R. Choudhary, R. Hida, M. Hamada, H. Futami, and T. Sekiya. Exploring context strategies in LLMs for discourse-aware machine translation. In **Findings of EMNLP 2025**, pp. 24382–24391, 2025.

[4] L. Wang, C. Lyu, T. Ji, Z. Zhang, D. Yu, S. Shi, and Z. Tu. Document-level machine translation with large language models. In **Proc. EMNLP 2023**, pp. 16646–16661, 2023.

[5] M. Wu, T. Vu, L. Qu, G. Foster, and G. Haffari. Adapting large language models for document-level machine translation, 2024.

[6] M. Ramos, P. Fernandes, S. Agrawal, and A. Martins. Multilingual contextualization of large language models for document-level machine translation. In **2nd COLM**, 2025.

[7] M. Finkelstein, D. Vilar, and M. Freitag. Introducing the NewsPaLM MBR and QE dataset: LLM-generated high-quality parallel data outperforms traditional web-crawled data. In **Proc. 9th WMT**, pp. 1355–1372, 2024.

[8] D. O'Brien, B. Malik, O. de Gibert, P. Chen, B. Haddow, and J. Tiedemann. Dochplt: A massively multilingual document-level translation dataset. In **Proc. 10th WMT**, pp. 286–300, 2025.

[9] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. Language-agnostic BERT sentence embedding. In **Proc. 68th ACL**, pp. 878–891, 2022.

[10] M. Cettolo, C. Girardi, and M. Federico. WIT3: Web inventory of transcribed and translated talks. In **Proc. 16th EAMT**, pp. 261–268, 2012.

[11] M. Cettolo, M. Federico, L. Bentivogli, J. Niehues, S. Stüker, K. Sudoh, K. Yoshino, and C. Federmann. Overview of the IWSLT 2017 evaluation campaign. In **Proc. 14th IWSLT**, pp. 2–14, 2017.

[12] Y. Jiang, T. Liu, S. Ma, D. Zhang, J. Yang, H. Huang, R. Sennrich, R. Cotterell, M. Sachan, and M. Zhou. BlonDe: An automatic evaluation metric for document-level machine translation. In **Proc. NAACL 2022**, pp. 1550–1565, 2022.

[13] M. Xu, L. Wang, D. Wong, H. Liu, L. Song, L. Chao, S. Shi, and Z. Tu. GuoFeng: A benchmark for zero pronoun recovery and translation. In **Proc. EMNLP 2022**, pp. 11266–11278, 2022.

[14] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In **Proc. 10th MTSummit**, pp. 79–86, 2005.

[15] Qwen Team. Qwen technical report, 2023.

[16] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. Manning. Stanza: A Python natural language processing toolkit for many human languages. In **Proc. 58th ACL: System Demonstrations**, 2020.

[17] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proc. 40th ACL**, pp. 311–318, 2002.

[18] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer. Multilingual denoising pre-training for neural machine translation. **TACL 2020**, pp. 726–742, 2020.

[19] V. Raunak, T. Kocmi, and M. Post. SLIDE: Reference-free evaluation for machine translation using a sliding document window. In **Proc. NAACL 2024**, pp. 205–211, 2024.

[20] R. Rei, J. C. de Souza, D. Alves, C. Zerva, A. Farinha, T. Glushkova, A. Lavie, L. Coheur, and A. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In **Proc. 7th WMT**, pp. 578–585. Association for Computational Linguistics, 2022.

[21] X. Lyu, J. Li, Z. Gong, and M. Zhang. Encouraging lexical translation consistency for document-level neural machine translation. In **Proc. EMNLP 2021**, pp. 3265–3277, 2021.

[22] T. Kocmi and C. Federmann. Large language models are state-of-the-art evaluators of translation quality. In **Proc. 24th EAMT**, 2023.

[23] OpenAI Team. Gpt-4 technical report, 2024.

[24] Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

[25] Kwon W., Li Z., Zhuang S., Sheng Y., Zheng L., Yu C., Gonzalez J., Zhang H., and Stoica I. Efficient memory management for large language model serving with page-dattention. In **Proc. 29th ACM SOSP**, 2023.

[26] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In **ICRL 2019**, 2019.

A　訓練・推論設定

モデルの訓練には SFT (supervised fine-tuning) を採用し, 損失はターゲット側に対してのみ算出する. 訓練および推論のプロンプトは以下に記載している.

Prompt for training and inference

<|im_start|>user
Context1: # if necessary
{source_lang}: {source_context1}
Context2: # if necessary
{source_lang}: {source_context2}
...
Translate the following source text from {source_lang} into {target_lang}.
{source_lang}: {source_text}.
{target_lang}: <|im_end|>
<|im_start|>assistant
{target_text}.<|im_end|>

訓練のハイパーパラメータは, Ramos ら [6] によって設定に従う. ただし, バッチサイズについては, Doc2Doc では 16, それ以外では 64 に調整した. すべての訓練プロセスは 8 個の NVIDIA H100 GPU を使用する. 推論は vLLM [25] を採用し, グリーディデコーディングを用い, 単一の H100 GPU 上で実施した. 詳細なパラメータ設定は表 2 に示す.

表 2: 訓練におけるハイパーパラメータ設定.

Batch size	16 (D2DFT) / 64 (others)
Number of Epochs	2
Learning rate	7×10^{-6}
LR Scheduler	cosine
Warmup Steps	125
Weight Decay	0.01
Optimizer	AdamW [26]
Adam β_1	0.9
Adam β_2	0.999
Adam ϵ	1×10^{-8}
Maximum Sequence Length	32,768

B　d-COMET の計算方式

チャンクトリプレット (src, hyp, ref) を 3 つの手法で構築する. なお, すべての戦略において, 連続するチャンク間で 50%のオーバーラップを維持する.

Algorithm 1: 二重境界マッチングによるチャンク対応付け

Input : Source chunks $\mathscr{C}$, Target sentences $\mathscr{T}$, λ, σ, Initial k, Δk
Output : Aligned lower bound indices $J = \{j_1, \ldots, j_{|\mathscr{C}|}\}$, $j \in \{1, .., |\mathscr{T}|\}$

Function $\texttt{SimwRP}(u, v)$**:**
  $s \leftarrow \text{CosSim}(u, v)$;
  $w \leftarrow \exp\left(-\left(\frac{\text{rp}(u)-\text{rp}(v)}{\sigma}\right)^2\right)$;
  return $s \times ((1-\lambda) + \lambda w)$;
Embedding for $c^{\text{end}}, c^{\text{first}} \in \mathscr{C}$ and $t \in \mathscr{T}$;
Initialize score matrix $\Phi \in \mathbb{R}^{|\mathscr{C}| \times |\mathscr{T}|}$;
while $J = \emptyset$ **do**
  for $i \leftarrow 1$ **to** $|\mathscr{C}|$ **do**
    for $j \leftarrow 1$ **to** $|\mathscr{T}|$ **do**
      $\Phi[i, j] \leftarrow \texttt{SimwRP}(c_i^{\text{end}}, t_j)$;
      Find top-k candidates $\{j_{i,1}, ..., j_{i,k}\}$;
      for $r \leftarrow 1$ **to** k **do**
        $\Phi[i, j] \leftarrow \Phi[i, j] + \texttt{SimwRP}(c_{i+1}^{\text{first}}, t_{j_{i,r}+1})$;
  Find $j_1, \ldots, j_{|\mathscr{C}|}$ maximizing $\sum \Phi[i, j_i]$;
    subject to $j_{i-1} \le j_i$ (Monotonicity);
    and $j_{|\mathscr{C}|} = |\mathscr{T}|$ (Boundary Condition);
if $J = \emptyset$ **then** $k \leftarrow \min(k + \Delta k, |\mathscr{T}|)$;
return J **or Failure**

最終的な d-COMET スコアは, これら 3 つの異なる手法から得られたスコアの平均として算出される.

- **Src–Ref Aligned Stacking**: ソースと参照訳をいずれかが 512 トークンに達するまで連結する. その後, 翻訳候補を参照訳の長さに合わせて分割・累積し, 全系列が 512 トークン未満となるよう調整する.
- **Segment Independent Stacking**: 系列間のアライメントを考慮せず, ソース, 翻訳候補, 参照訳の各セグメントを, それぞれ最大容量の 512 トークンに達するまで独立して連結する。
- **Token Independent Stacking**: セグメント境界やアライメントを無視し, ソース, 翻訳候補, 参照訳の各系列に対して, 個別に 512 トークンのスライディングウィンドウを適用する。