機械翻訳における多様な候補の多段階洗練と生成的統合

河田怜治　後藤功雄　二宮崇
愛媛大学
{kawata@ai.cs., goto.isao.fn@, ninomiya.takashi.mk@}ehime-u.ac.jp

概要

大規模言語モデルによる翻訳は流暢だが，幻覚や微細な誤訳等の課題が残る．推論時の事後修正により改善する手法があるが，単一モデルによる修正はバイアスに弱く，複数出力の統合手法は入力品質に依存する．そこで本研究では，多様なモデルによる候補生成と個別洗練を経て統合を行う 4 段階の枠組みを提案する．WMT データセットを用いた実験の結果，提案手法は既存手法と比較して高い xCOMET スコアを記録した．

1　はじめに

大規模言語モデル（Large Language Model; LLM）の急速な発展により，機械翻訳の流暢さは飛躍的に向上した．文脈を考慮した自然な訳の生成が可能になった一方で，原文に含まれない情報を生成する「幻覚（Hallucination）」や，文脈の微細なニュアンスを取り違える意味的な誤りは依然として解決するべき課題である．こうした誤りを修正するために，モデル自体のパラメータを更新することは膨大な計算コストを要するため，近年では生成された翻訳結果を事後的に改善するアプローチが数多く提案されている [1–6]．既存手法は「単一候補の洗練」と「複数候補の活用」に大別される．前者は Iterative Translation Refinement [1] 等が代表的で，DUAL-REFLECT [2] は逆翻訳を用いて整合性を検証するが，単一モデルのバイアスや知識不足に弱い．一方，後者は候補選択や統合を行う．最近提案された Mixture-of-Agents（MoA） [5] は生成的統合を LLM のベンチマークで評価して有効性を確認したが，翻訳タスクを対象としたものではない．

そこで本研究では，MoA を翻訳に応用し，複数の異なる LLM を用いた生成的統合と逆翻訳による検証・修正に基づく機械翻訳手法を提案する．多様なモデルによる候補生成と個別の洗練を経てから統合を行うことで，単一モデルゆえの制約の解消を目指す．実験の結果，提案手法は xCOMET スコアにおいて高い性能を記録し，本アプローチの有効性を確認した．

2　関連研究

翻訳文の事後修正　LLM の翻訳品質を向上させるため，生成結果を反復的に修正する手法が注目されている．Chen ら [1] は LLM に自己修正を行わせる Iterative Translation Refinement を，Vernikos ら [6] は小規模モデルを用いて出力を書き直す LMCor を提案し，ファインチューニングなしでの性能向上を示した．本研究では，これらの手法と同様に LLM を用いた出力の書き直しを採用するが，これを最終的な修正ではなく，後段の逆翻訳や統合を行う前の予備的な改善として位置づけている．

逆翻訳を用いた品質検証　逆翻訳は原文の意味保持を検証する有効な手段である．DUAL-REFLECT [2] は，翻訳文を逆翻訳して原文と比較し，その差異をフィードバックとして修正を行う．しかし，この手法は生成・逆翻訳・評価が単一モデル内で完結するため，モデル固有の知識バイアスにより誤りを見落とすリスクがある．本研究では，この検証プロセスを複数モデルの候補に対して適用することで単一モデルの盲点を克服し，より信頼性の高い修正を実現する．

翻訳候補のアンサンブルと統合　不確実性低減のため，複数の翻訳候補を活用する研究も盛んである．Wang ら [3] は MBR Decoding による候補選択の有効性を示し，Vernikos ら [4] は QE-Fusion による部分的統合を提案した．GenerateBest [3] のような再生成手法があるが，単一モデル由来のため入力の多様性に欠ける．LLM のベンチマークを対象に，複数のモデルを多段に配置し，前段の出力を参照して回答精度を高める Mixture-of-Agents (MoA) [5] が高い性能を示しているが，翻訳タスクを対象としたものではない．本研究では，翻訳タスクに異種モデルによる多様性確保と逆翻訳による事前検証を導入する．

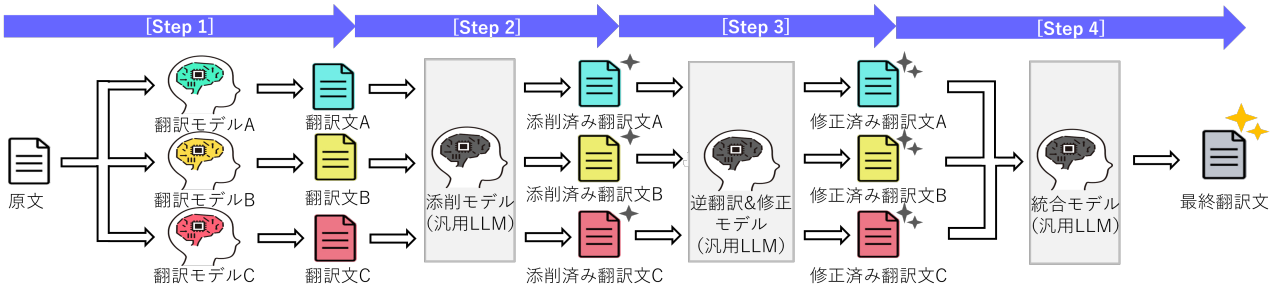

図 1: 提案手法の概要図

表 1: 実験に使用したデータセットの文数

言語対	WMT23	WMT24
英 → 日 (En-Ja)	2,074	997
英 → 中 (En-Zh)	2,074	997
英 → 独 (En-De)	557	997

表 2: 各段階で使用したモデル

段階	モデル
初期候補生成	GPT-4o [1] [7]
	XALMA [2] [8]
	Qwen3-32B [3] [9]
添削・検証・統合	GPT-4o

3 提案手法

本研究では，単一モデルが抱える幻覚や局所的な誤りを克服し，多様な知識を含む翻訳候補を統合することで高品質な翻訳を実現するため，個々の翻訳候補文の反復修正と生成的統合を組み合わせた 4 段階のフレームワークを提案する．本手法の全体像を図 1 に示す．プロセスは以下の 4 つのステップで構成される．

1. 複数モデルによる多様な候補生成
2. 翻訳候補の予備的改善
3. 逆翻訳による検証
4. 生成的統合

各ステップにおいて実験に使用したプロンプトは，付録 A に記載している．

以下に，各ステップの詳細を述べる．

Step 1: 複数モデルによる多様な候補生成　最初の段階として，入力文に対して複数の異なる翻訳モデル（$M_1, M_2, \cdots, M_n$）を用いて初期翻訳候補を生成する．既存の統合手法の多くは，単一のモデルからサンプリングパラメータ（Temperature 等）を変えて複数の候補を得るが，モデル固有の知識バイアスや苦手な表現は共通しているため，本質的な多様性は得られにくい．本手法では，学習履歴の異なる複数の LLM を採用することで，語彙選択や構文構造における多様性を確保する．これにより，後段の統合プロセスにおいて，相互補完的な情報が得られる可能性を高める．

Step 2: 翻訳候補の予備的改善　生成された初期翻訳には，文法的な誤りや不自然な言い回し，冗長な表現などが含まれる場合がある．これらは後段の逆翻訳と修正のプロセスにおいてノイズとなる可能性がある．そこで本ステップでは，これらを修正することを目的として，LLM に「翻訳の添削者」としての役割を与える．原文と初期翻訳候補文を同時に入力し，翻訳品質の向上を指示するプロンプトを用いる．この際，翻訳がすでに自然かつ正確である場合には，一切の変更を加えずにそのまま出力するよう明示的に指示する．これにより，各翻訳候補文の流暢さと文法的な正しさを改善し，基礎品質を底上げする．

Step 3: 逆翻訳による検証　表層的な流暢さが確保されたとしても，原文の意味を正しく反映していない意味的誤りが残存している可能性がある．このような誤りには，入力文に含まれない情報が付加される「幻覚」だけでなく，原文中の情報が十分に反映されていない翻訳不足も含まれる．特に後者は，機械翻訳において頻繁に観察される問題である．これらの意味的誤りを検出・修正するため，本ステップでは逆翻訳を用いた検証を反復的に行う．具体的には，Step 2 で得られた候補文を翻訳元言語へ逆翻訳し，元の入力文と比較する．逆翻訳文において原文中の情報が欠落している，あるいは内容が大きく乖離している場合，その候補文には翻訳不足や意味的な歪みが含まれていると判断できる．この差異をフィードバックとしてモデルに提示し，候補文を書き直させる．この逆翻訳と原文との比較・修正は，予め設定したターン数が終了するまで行う．

1) https://platform.openai.com/docs/models/gpt-4o
2) https://huggingface.co/haoranxu/X-ALMA
3) https://huggingface.co/Qwen/Qwen3-32B

表 3: WMT23 および WMT24 における翻訳性能比較

手法	モデル	WMT23						WMT24					
		英日		英中		英独		英日		英中		英独	
		xCOMET	BLEU	xCOMET	BLEU	xCOMET	BLEU	xCOMET	BLEU	xCOMET	BLEU	xCOMET	BLEU
Step 1	GPT-4o	89.07	24.09	87.63	48.24	87.83	**43.69**	83.02	29.75	79.82	**43.51**	91.06	**34.10**
	XALMA	87.49	22.54	86.49	47.72	86.63	38.32	77.82	24.63	76.13	40.12	90.40	32.02
	Qwen3-32B	87.41	18.84	85.21	38.29	81.95	27.14	76.37	22.15	76.53	32.91	87.98	25.86
Step 1+2	GPT-4o(GPT-4o)	88.82	24.40	87.51	48.18	87.92	43.65	82.99	29.70	79.57	43.33	91.08	34.06
	GPT-4o(XALMA)	88.57	24.07	87.54	**48.98**	87.68	40.25	79.96	26.04	77.99	42.26	91.04	33.11
	GPT-4o(Qwen3-32B)	88.62	23.00	86.70	40.91	86.50	34.10	79.40	23.68	78.72	35.00	90.36	28.80
DUAL-REFLECT	GPT-4o	87.08	20.97	84.96	40.97	86.94	43.03	79.94	26.83	76.19	38.24	89.66	31.02
GenerateBest	GPT-4o	88.59	24.15	86.95	48.86	87.45	42.43	80.79	27.79	78.16	41.71	90.75	33.29
Step 1+2+4	GPT-4o $\left\{\begin{smallmatrix}\text{GPT-4o,}\\ \text{XALMA,}\\ \text{Qwen3-32B}\end{smallmatrix}\right\}$	**90.44**	**24.90**	**88.04**	47.04	88.16	42.56	**83.46**	**29.99**	**80.32**	42.70	**91.61**	33.99
Step 1+2+3+4	GPT-4o $\left\{\begin{smallmatrix}\text{GPT-4o,}\\ \text{XALMA,}\\ \text{Qwen3-32B}\end{smallmatrix}\right\}$	89.60	24.69	87.78	46.94	**88.34**	40.99	83.22	29.92	79.72	42.38	91.08	33.00

手法列の Step 1 は 0-shot 翻訳，Step 1+2 は添削，Step 1+2+4 は添削済み候補文に対して逆翻訳による検証を省いた生成的統合，Step 1+2+3+4 は全工程を指す．DUAL-REFLECT，GenerateBest は比較用の既存手法である．モデル列の括弧内は，統合や添削の入力となる候補の生成元モデルを表す．各言語対において**太字**は最高値，下線は次点を示す．

Step 4: 生成的統合　最終段階として，Step 3 で意味的な正確さを向上させた複数の候補文を統合し，最終的な出力を生成する．本手法では，LLM を用いた生成的統合を採用する．LLM のプロンプトに検証済みの複数の候補文を入力し，それぞれの長所（正確な用語や自然な構文など）を組み合わせるように指示することで，単一の候補では到達できない，高い頑健性と流暢さを兼ね備えた翻訳文の生成を目指す．

4　実験

4.1　実験設定

データセット　評価用データセットとして，WMT23 [10] および WMT24 [11] のテストセットを使用した．言語対は，言語構造が大きく異なる 3 方向 (英-日, 英-中, 英-独) を採用した．各データセットの文数を表 1 に示す．

使用モデル　本実験の各段階で使用したモデルを表 2 に示す．初期翻訳候補の生成には，多様性を確保するために特徴の異なる 3 つのモデルを採用し，その後の添削・検証・統合には，指示追従性に優れた GPT-4o を使用した．

比較手法　提案手法の有効性を検証するため，既存手法，および提案手法の各段階に対応する計 6 つの設定で比較を行った．

- **Step 1 (0-shot 翻訳)**: 初期候補生成のみを行ったベースライン．GPT-4o, XALMA, Qwen3-32B の 3 モデルそれぞれの生成結果を評価した．
- **Step 1+2 (0-shot + 添削)**: Step 1 の生成結果に対し，GPT-4o で Step 2 の添削を行ったもの．
- **DUAL-REFLECT [2]**: 単一モデルで初期候補を生成し，逆翻訳と修正を繰り返す既存手法．
- **GenerateBest [3]**: 単一モデルで複数の翻訳候補文を生成し，それらを入力として新たな訳文を生成する既存手法．
- **Step 1+2+4**: Step 3（逆翻訳による検証）を省いた構成．
- **Step 1+2+3+4**: 添削済みの候補文に対して逆翻訳と修正のターンを繰り返し，その後各候補文を参考にして新たな翻訳文を生成する完全な構成．本実験では，逆翻訳と修正を 3 ターン行った．

評価指標　本研究では，翻訳品質を多角的に評価するために，xCOMET [12] と BLEU [13] の 2 つの指標を採用した．xCOMET は，事前学習済み言語モデルを用いてソース文，生成文，参照訳の意味的な類似度を評価する指標であり，BLEU は，生成された翻訳文と参照訳との間で n-gram の一致率を計算する指標である．

4.2　実験結果

WMT23 および WMT24 データセットにおける実験結果を表 3 に示す．なお，WMT24 の英独 (En-De) 翻訳タスクでは，評価用に 2 つの参照訳が提供されているため，本実験ではそれぞれの参照訳に対して

表 4: Step3 でのターンごとの評価（WMT23 および WMT24）

モデル	ターン	WMT23						WMT24					
		英日		英中		英独		英日		英中		英独	
		xCOMET	BLEU	xCOMET	BLEU	xCOMET	BLEU	xCOMET	BLEU	xCOMET	BLEU	xCOMET	BLEU
GPT-4o (GPT-4o)	0	88.82	24.40	87.51	48.18	87.92	43.65	82.99	29.70	79.57	43.33	91.08	34.06
	1	89.30	24.29	87.26	48.04	87.99	43.13	82.83	29.81	79.02	43.10	90.68	33.53
	2	89.29	24.42	87.46	48.12	88.03	42.89	82.78	29.65	79.03	43.26	90.69	33.43
	3	89.04	24.35	87.36	47.95	88.15	42.69	82.64	29.52	78.95	43.20	90.57	33.24
GPT-4o (XALMA)	0	88.57	24.07	87.54	48.98	87.68	40.25	79.96	26.04	77.99	42.26	91.04	33.11
	1	89.15	24.41	87.47	49.11	88.01	41.85	81.09	28.18	78.45	44.03	90.77	33.83
	2	89.15	24.52	87.51	49.17	88.07	41.87	81.25	28.74	78.53	43.94	90.71	33.84
	3	89.09	24.28	87.55	49.17	88.01	41.90	81.26	28.76	78.55	43.91	90.57	33.56
GPT-4o (Qwen3-32B)	0	88.62	23.00	86.70	40.91	86.50	34.10	79.40	23.68	78.72	35.00	90.36	28.80
	1	89.22	23.82	86.89	45.08	88.10	39.43	81.68	28.42	79.03	40.16	90.48	32.05
	2	89.23	23.84	86.82	45.56	88.05	39.64	81.83	28.65	79.01	40.68	90.47	32.32
	3	89.13	23.97	86.89	45.78	88.10	39.75	81.67	28.81	78.87	40.86	90.59	32.32

モデル列の括弧内は，入力となる添削済み候補の生成元モデルを表す．

算出されたスコアの平均値を採用した．

意味的類似度指標の xCOMET において，複数モデルの出力の生成的統合（Step 1+2+4）および，それに逆翻訳を取り入れた逆翻訳による検証+生成的統合（Step 1+2+3+4）は，0-shot 翻訳や，既存手法と比較して，一貫して高い性能を達成した．具体的には，WMT23 の英日翻訳（En-Ja）において，生成的統合は 0-shot の GPT-4o と比較して xCOMET スコアで+1.37 ポイント向上した．英独（En-De）においては，逆翻訳による検証+生成的統合が全ての手法を上回り，最高スコア（88.34）を記録している．これに対し，n-gram 一致率に基づく BLEU スコアでは，0-shot 翻訳が最も高い値を示す傾向が見られた．WMT24 においても同様に，生成的統合および逆翻訳による検証+生成的統合の xCOMET における有効性が次のように確認された．英日（En-Ja）においては，生成的統合の xCOMET スコア 83.46 が 0-shot 翻訳（83.02），DUAL-REFLECT（79.94），GenerateBest（80.79）を上回った．英中（En-Zh）および英独（En-De）においても，生成的統合が最も高い xCOMET スコア（それぞれ 80.32, 91.61）を記録している．この「検証プロセスの有無による性能差」については，次節の分析において詳述する．

4.3 分析

本節では，Step 3 における逆翻訳検証・修正プロセスが翻訳品質に与える影響について分析する．表 4 に，Step 3 の各ターンにおける各候補文の品質推移を示す．結果を確認すると，多くのケースでターンを経るごとに，各候補文の xCOMET スコアは Step 2（表 4 のターン 0）と比較して向上，あるいは高い水準を維持していることがわかる．例えば，WMT23 英日における GPT-4o 由来の候補文のスコアは，Step 2 の 88.82 から，Step 3 のターン 1 で 89.30 へと上昇している．これは，逆翻訳を用いた自己検証と修正が，個々の翻訳候補の品質改善に対して有効に機能していることを示唆している．

しかし，最終的な統合結果（表 3）において，これら検証済み候補文を用いた「Step 1+2+3+4（逆翻訳による検証+生成的統合）」は，検証を経ずに統合した「Step 1+2+4（生成的統合）」と比較して，WMT23 英独を除く多くの設定でスコアが低下する結果となった．個々の候補の品質が向上しているにもかかわらず統合後の性能が伸び悩んだ要因として，候補間の「多様性の喪失」が考えられる．MoA のような生成的統合手法は，入力される複数の候補が持つ異なる知識や表現の多様性を活用することで，単一モデルでは得られない高品質な出力を生成する．しかし，本実験では，Step 3 では全ての候補に対して同一のモデルを用いて逆翻訳および修正を行った．これにより，元々は異なるモデルによって生成された多様な候補が，検証に用いた特定モデルの知識や選好へと寄せられ，結果として表現の均質化を招いたと推察される．したがって，今後は候補の品質を担保しつつ，多様性を維持するために，検証モデル自体も多様化させるなどの検討が必要である．

5 おわりに

本研究では，複数 LLM の生成的統合と逆翻訳検証に基づく機械翻訳手法を提案した．WMT テストセットでの実験の結果，提案手法は既存手法を上回る xCOMET スコアを達成した．今後は，検証モデルの多様化や公開モデルの活用が課題である．

謝辞

本研究は国立研究開発法人情報通信研究機構の委託研究（課題番号：225）および JSPS 科研費 JP24K15071 および大学発新産業創出基金事業スタートアップ・エコシステム共創プログラム JPMJSF2316 の助成を受けたものです.

参考文献

[1] Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. Iterative Translation Refinement with Large Language Models. In **Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)**, pp. 181–190, 2024.

[2] Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. DUAL-REFLECT: Enhancing Large Language Models for Reflective Translation through Dual Learning Feedback Mechanisms. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 693–704, 2024.

[3] António Farinhas, José de Souza, and Andre Martins. An Empirical Study of Translation Hypothesis Ensembling with Large Language Models. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 11956–11970, 2023.

[4] Giorgos Vernikos and Andrei Popescu-Belis. Don't Rank, Combine! Combining Machine Translation Hypotheses Using Quality Estimation. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 12087–12105, 2024.

[5] Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Y Zou. Mixture-of-Agents Enhances Large Language Model Capabilities. In **International Conference on Learning Representations**, pp. 33944–33963, 2025.

[6] Giorgos Vernikos, Arthur Brazinskas, Jakub Adamek, Jonathan Mallinson, Aliaksei Severyn, and Eric Malmi. Small Language Models Improve Giants by Rewriting Their Outputs. In **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2703–2718, 2024.

[7] OpenAI. GPT-4o system card. **arXiv preprint arXiv:2410.21276**, 2024.

[8] Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. X-ALMA: Plug & play modules and adaptive rejection for quality translation at scale. In **The Thirteenth International Conference on Learning Representations**, 2025.

[9] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report. **arXiv preprint arXiv:2505.09388**, 2025.

[10] Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, Mariya Shmatova, and Jun Suzuki. Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet. In **Proceedings of the Eighth Conference on Machine Translation**, pp. 1–42, 2023.

[11] Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In **Proceedings of the Ninth Conference on Machine Translation**, pp. 1–46, 2024.

[12] 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. **Transactions of the Association for Computational Linguistics**, Vol. 12, pp. 979–995, 2024.

[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, 2002.

付録

A　入力プロンプト

本実験で提案手法の各 Step で使用したプロンプトを図 2 および図 3 に示す．図 2 は候補生成，添削，統合（Step 1, 2, 4），図 3 は逆翻訳による検証プロセス（Step 3）のプロンプトである．

Step 1: 初期候補生成

System:
You are a professional translator. Translate the following `{src_lang}` text into natural-sounding `{tgt_lang}`.

User:
`{Input Source Sentence}`

Step 2: 翻訳候補の予備的改善

System:
You are a professional translation reviewer.

User:
Given the `{src_lang}` source sentence and its `{tgt_lang}` translation, rewrite the `{tgt_lang}` sentence to make it more natural and accurate. If the original translation is already natural and accurate, output it as is without any change. Output only the final `{tgt_lang}` sentence.

 # `{src_lang}` source
`{{src_lang} Source Sentence}`
 # Original `{tgt_lang}` translation
`{Original {tgt_lang} Translation}`

Step 4: 生成的統合

System:
You are an expert translator specialized in `{src_lang}` to `{tgt_lang}` translation.

User:
Please provide the best possible `{tgt_lang}` translation for the following `{src_lang}` sentence. Three existing translations are provided for reference.

[{src_lang} Original] `{Source Sentence}`
[Reference Translation 1] `{Candidate 1}`
[Reference Translation 2] `{Candidate 2}`
[Reference Translation 3] `{Candidate 3}`
[Your Best Translation]

図 2: 実験に使用したプロンプト（Step 1, 2, 4）

Step 3: 逆翻訳による検証と修正

[Step 3-1: 逆翻訳]
System: You are a professional translator specialized in `{tgt_lang}`-`{src_lang}` translation. You will receive 3 `{tgt_lang}` candidate translations derived from the same source. Your task is to produce back-translations to verify the semantic accuracy of these candidates.

User: Please back-translate each of the 3 `{tgt_lang}` candidates into `{src_lang}`.
Important Instructions:
- **Prioritize Semantic Fidelity**: Ensure the back-translation accurately reflects the meaning, nuance, and tone of the `{tgt_lang}` candidate.
- **Do NOT Repair**: If a `{tgt_lang}` candidate is ambiguous, unnatural, or grammatically incorrect, reflect that flaws in the back-translation (do not "fix" it to make it sound better in `{src_lang}`).
- **Independence**: Treat each candidate independently.

`[{tgt_lang} Candidate 1]{Candidate Text 1}` ...

[Step 3-2: 比較＆検証]
System: You are a highly skilled bilingual editor specialized in `{src_lang}`-`{tgt_lang}` translation refinement. Your goal is to ensure semantic accuracy and naturalness.

User: You are given: 1. The original source. 2. Several back-translations. 3. Several candidate translations. **Your task:**
- Detect meaning deviations between the source and the back-translations.
- Use all `{tgt_lang}` candidates as references.
- Produce ONE improved `{tgt_lang}` translation.

Input:
`[{src_lang} Source]{Source Sentence},`
`[Back-Translations]{Back-Translation 1...},`
`[{tgt_lang} Candidates]{Candidate 1...}`

図 3: 実験に使用したプロンプト（Step 3）