ABCI 3.0開発加速利用 （2025年度）成果概要（公開用）

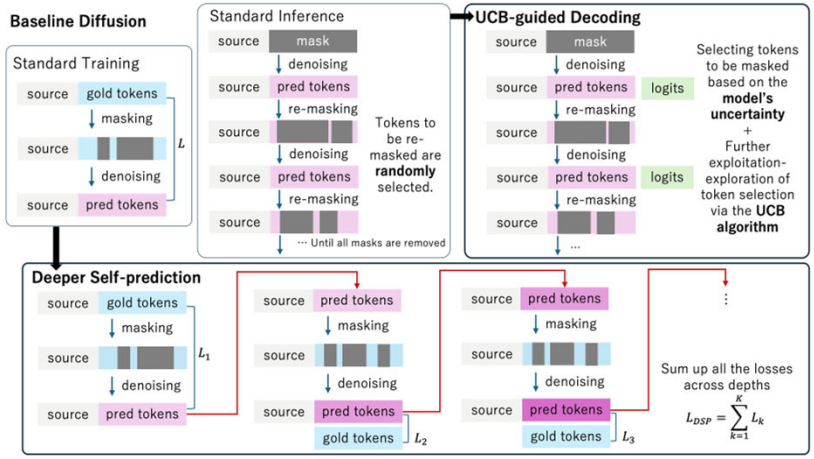
課題名： AIセーフティ強化に関する研究開発「マルチモーダルAIモデル開発用データのセーフティ管理技術の研究開発」	実施時期：2025年4月～2026年3月 所属機関名：産業技術総合研究所 代表者氏名：浅田　真生

成果概要：

AIによるテキスト生成では、学習時と推論時の乖離により、不正確または不適切な出力が生じ得る。本研究では、この課題に対処し生成過程のセーフティ向上を目的として、新たな生成フレームワークを提案した。本手法は、生成結果を反復的に修正する学習機構と、生成時に再検討箇所を選択する推論手法を統合したものである。これにより誤りの蓄積を抑制し、より適切で一貫性のある出力を実現する。3種のタスクで評価を行い、既存手法と比較して事実整合性および流暢性の向上を確認した。

成果のポイント：

① 自己修正に基づく学習手法の開発
従来のテキスト生成モデルでは、訓練時と推論時の乖離が不正確な出力の一因となっていた。本研究では、生成途中の予測を次段の入力として用い、誤りを段階的に修正する学習手法を開発した。これにより、生成過程に近い条件での学習が可能となり、不確実性の把握能力も向上した。
② 見直し箇所選択のための推論アルゴリズムの開発
反復的な生成過程において、見直し箇所の選択は品質に大きく影響する。従来手法では誤った確信に基づく出力に陥る可能性があった。本研究では、探索と活用の観点から選択手法を設計し、誤りの見逃しを抑制しつつ安定した改善を実現した。推論時の追加計算コストは限定的である。
③ 複数タスク・モデル規模における性能向上の検証
複数のベンチマークで評価を行い、すべてのタスクで既存手法を上回る性能を確認した。特に事実整合性および正解率の改善が見られた。異なるモデル規模でも同様の傾向が確認され、推論時の追加コストはほぼ生じない。



成果についてより詳細な情報を提供しているWebページ、発表論文などの情報：

【発表論文】

- Masaki Asada, Makoto Miwa. “Principled Self-Correction in Discrete Diffusion: A UCB-Guided Framework for Text Generation.” Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026), Volume 1: Long Papers, pages 6678–6692, March 2026.
- 浅田 真生、三輪 誠　「自己修正学習とUCBデコーディングによる離散拡散テキスト生成」　言語処理学会 第32回年次大会 発表論文集 pages 4314-4319

【コード公開先】 https://github.com/aistairc/UCB-DSP-DiffusionLM