棄却サンプリングを用いたLLMによる制限自然言語生成の評価

伊良皆慧斗　山田孝治
琉球大学　{e225746,koji}@ie.u-ryukyu.ac.jp

概要

自然言語は曖昧性を持つため、構造化テキストや形式言語のような解析がしやすい出力が求められる。一方で、出力の効率性と忠実性にはトレードオフがあることが知られている。制限自然言語は自然言語に近いため、このトレードオフを乗り越えられる可能性がある。しかし、LLMの制限自然言語の能力はよく知られていない。本研究は、LLMによる制限自然言語生成をハード制御の制御可能なテキスト生成として定式化し、GUARDによる保証された制限自然言語生成を検証する。複数の制限自然言語及びプロンプトで受理率、忠実性及び多様性の評価を行い、受理率が改善されることが示された。

1　はじめに

LLMは様々な自然言語処理タスクで成果を上げており[1]、ChatGPTのような対話型の利用から、AIエージェントのような自律型の利用がされるなど、様々なシステムに組み込まれている。LLMは人と同等程度に自然な文章の生成をすることができる。一方で自然言語の文は2つ以上の意味で解釈される曖昧性を持つ場合がある[2]。自然言語の曖昧性は下流のシステムの誤作動を引き起こす可能性がある。自然言語の文を生成させ利用する場合には、適切な処理が必要となる。そのため、実際にLLMをシステムに組み込む際には、JSON形式などの構造化テキストや形式言語の文など、機械的に処理が容易な出力をすることが必要とされる。与えられた構造や文法の制約を完全に守りつつ出力する手法として、制約付きデコーディングや棄却サンプリングが挙げられる。しかし、制約付きデコーディングはLLMの出力の分布を歪ませる忠実性の問題、棄却サンプリングは満たすのが難しい制約に対して多くの再生成が求められる効率性の問題がある。制約を満たす生成において、この2つにトレードオフがあることが指摘されている[3]。このトレードオフが生じる原因の1つとして、LLMが大量の自然言語文で学習されているのに対し、構造化テキストなどの文は、自然言語とは異なる構造の制約を課すことが挙げられる。よってトレードオフを乗り越えるためには、より自然言語に近くかつ曖昧性や複雑さの低い文の生成をすることが必要であり、トレードオフの解消はLLMを組み込む様々なシステムにおいて有用であると考えられる。制限自然言語は、自然言語をベースとした人工言語であり曖昧性や複雑さが軽減されている。そのため、制限自然言語生成は構造化テキスト生成などよりも効率が良く生成ができ、また忠実性を歪ませない生成が比較的容易であると考えられる。本研究では、LLMモデルにプロンプトを与え、制限自然言語文の出力を行い、その効率性や忠実性及び多様性の評価する。また、その結果をCNLの制約のタイプやPENS分類体系に基づき考察する。

2　関連研究

制限自然言語 (Controlled Natural Language; CNL) は、ある特定の自然言語に基づいて構築された言語で、語彙・構文・意味論に関してより制限的である一方、その自然な性質の大部分を保持しているものである[4]。制限自然言語を分類する体系として、正確さ (Precision)・表現力 (Expressiveness)・自然さ (Naturalness)・単純さ (Simplicity) の5段階評価を行うPENS分類体系 (PENS Classification Scheme) がある[4]。各評価項目のより厳しい基準を満たすほど、その数値は高くなる。英語をはじめとする自然言語はP1E5N5S1に分類される。正確さや単純さが最低評価であるものの、表現力や自然さは最高評価である。一方で命題論理は逆に表現力や自然さが最低評価であるものの、正確さや単純さは最高評価である。PENS分類体系は形式言語や自然言語を基準として、その中間の性質の持つ制限自然言語を分類する。特にN≥3であり、自然言語をベースにしている言語を制限自然言語と呼ぶ[4]。

制限自然言語の例として、英語ベースのAttempto Controlled English(ACE)や日本語ベースの「やさし

い日本語」がある。Attempto Controlled English は英語をベースとする制限自然言語である [5]。ACE は ACE Parsing Engine(APE)[1] を使って、談話構造表現などの形式言語や Web Ontology Language (OWL) へと翻訳することができる。そのため形式言語に近い性質を持った制限自然言語であるといえる。PENS 分類体系による分類は、P4E3N4S3 である [4]。

やさしい日本語は日本語をベースとする制限自然言語である。日本語非母語話者等が理解しやすくするために提案されているものである [6, 7]。やさしい日本語は平易化された日本語としても考えられ、ACE のような体系的仕様が定められているわけではない。一方で、やさしい日本語の計量的な分析によれば、品詞構成や語彙難易度等の言語的特徴が異なることが示されており [8]、また、やさしい日本語のガイドラインが文化庁より公開されている [9]。やさしい日本語の評価ツールとして、やさにちチェッカー [10] がある。これは、語彙、漢字、硬さ、長さ、文法の 5 つの観点で 5 段階評価を行う。PENS 分類体系による分類に従ってやさしい日本語を分類すれば、P2E5N4S2 に当てはまる。

近年、LLM に制限自然言語を生成させる研究が行われている [11, 12]。知識グラフの質問応答においては、ユーザが入力した自然言語の文から制限自然言語の一種である SQUALL を LLM を用いて生成を行う研究がされており、論理形式に直接変換をする場合よりも高い性能を得ている [11] 。また、ユーザが入力した自然言語の文を LLM によって、制限自然言語の一つである InsurLE に変換することで、LLM の幻覚を回避しつつ LLM の持つ構文解析能力を利用する手法が提案されている [12]。シンプルさに関するスタイル変換において、やさしい日本語コーパスの MATCHA[6, 7] を使用して、プロンプトによる LLM のスタイル変換性能を評価した研究がある [13]。このように自然言語入力をし、何かしらの特定のタスクをこなすために、LLM に自然言語を制限自然言語に変換させる LLM による制限自然言語生成研究が実施されている。しかし、LLM の出力を制限自然言語に制御する方法について焦点を当てた研究は調査した範囲では見つからなかった。

言語モデルが、与えられた制約を満たすようにテキスト生成をするタスクは制御可能なテキスト生成 (Controlled Text Generation CTG) と呼ばれる。制御可能なテキスト生成は大きく 2 つに分類することができ、それぞれハード制御とソフト制御に分けられる [14]。ハード制御は生成された文章の構造や語彙といった特定の要素に焦点を当てたものであり、ソフト制御は感情、文体、話題といった文章の抽象的な言語的属性に焦点を当てる。ハード制御はソフト制御に比べ、難しいタスクであると考えられている [14]。語彙や構文の制御をする制限自然言語生成はハード制御に分類される。制限自然言語生成を制御可能なテキスト生成の個別化されたタスクとして位置付けることは、当該分野の手法を応用する上で重要となる。また、やさしい日本語など平易性を目的としている制限自然言語もあり、スタイル変換などの分野の手法も制限自然言語生成において重要であると考えられる。

制御可能なテキスト生成の分野において、保証された生成、すなわち生成されるすべての出力が制約を厳密に満たすことを保証することに主眼を置いた手法として、GUARD フレームワークが提案されている [15]。この手法は、任意の論理制約に対して適用な可能である。そのため、出力された文が制限自然言語かどうかを制約条件とすることができる。GUARD は、プロンプト設計や追加学習によって制約を満たしやすくしたモデルを提案モデルとして用いる。提案モデルが生成した候補文に対して制約判定に基づく棄却サンプリングを行い、制約を満たす文のみを採択することで生成文を得る。提案モデルと目標分布の乖離度合いは、忠実性 (目標とする分布 gold filtered model との KL ダイバージェンス) と効率性 (提案分布からのサンプリング時の受理率) の両者の上界となる [15]。

制約付きデコーディングは構造化テキストの分野で使われている。制約付きデコーディングのメリットとして、一定の追加の計算コストで確実に制約を満たすことができる。一方でトークンの確率分布を歪めるため、忠実性が悪くなってしまう [3]。また一般にトークン単位での判定をする手法であるため、制限自然言語の文法に従うかを文単位ではなくトークン単位で調べるための方法が必要となる。既存の制限自然言語ツールは文単位で判定することが多いため、制約付きデコーディングへの活用が難しい。また、トークンの確率を触ることなどができないブラックボックスな LLM では、制約付きデコーディングの手法は適用できない。

1） https://github.com/Attempto/APE/

3 提案手法

GUARD を用いた制限自然言語文章の生成、特に学習手法を使わずプロンプトのみで制約条件を満たしやすくした LLM を提案モデルとし、制約条件を満たすかによる棄却サンプリングよる生成を提案する。この手法では、ブラックボックス LLM など重みの更新やトークンの確率を操作できない場合でも適用することができる利点がある。

4 評価実験

LLM に文の生成を指示し生成された文のうち、最初の一文を評価対象とした。制限自然言語生成の誘導をする場合には制限自然言語を、Baseline ではそのベースとなる自然言語の文の生成を指示した。効率性と忠実性について評価を実施し考察を行った。

4.1 プロンプト

制約認識プロンプトは、制約を満たすように出力を誘導するプロンプトである。制約認識プロンプトとして、Zero-shot、One-shot、Few-shot(5 件)、例文の代わりに制限自然言語の言語仕様をプロンプトに含めた Zero-Shot(spec) の 4 種類を与え、制限自然言語生成を指示した。Baseline ではベースとなる自然言語の生成を指示した。

4.2 制限自然言語

日本語ベースの制限自然言語として、やさしい日本語。英語ベースの制限自然言語として、Attempto Controlled English を用いる。与えられた文がやさしい日本語であるかの判定には、やさにちチェッカーにて、全ての評価項目が 4 以上であるという閾値による判定の基準にする。ACE では、Clex 辞書を使用し未知語を含む場合も自動的に品詞を推測する設定で、APE 受理するを基準とする。

4.3 評価指標

効率性の評価として受理率 $AR_{a'}$、忠実性の評価として KL ダイバージェンス $\mathrm{KL}(g \,\|\, g')$ を求める。また、先行研究 [15] より，$\mathrm{KL}(g \,\|\, a') = \mathrm{KL}(g \,\|\, g') - \log AR_{a'}$ が成り立つため、総合的な評価としてこの値の推定した。KL ダイバージェンスは、ベースラインの受理率が低いため、自己正規化重点サンプリング (SNIS) による推定を行った。推定の妥当性を見るため、ユニークな文の数および、有効サンプルサイズ (ESS) を求めた。

5 実験結果

5.1 受理率と多様性の評価

表 1 ACE の受理率とユニークな文数

手法	生成数 n	受理数 n_{ac}	受理率 (%)	生成数 (unique.)	受理数 (uniq.)
baseline	6812	1	0.14	95	1
0-shot	6812	2565	37.6	609	196
1-shot	6811	6811	100	2	2
5-shot	6811	6811	100	3	3
spec	6811	6566	96.4	235	206

表 2 やさしい日本語の受理率とユニークな文数

手法	生成数 n	受理数 n_{ac}	受理率 (%)	生成数 (uniq.)	受理数 (uniq.)
baseline	6811	1019	14.9	4910	748
0-shot	6811	2232	32.7	4324	1155
1-shot	6811	3271	48.0	6108	2788
5-shot	6811	5667	83.2	3048	2312
spec	6811	3290	48.3	1977	1117

5.2 効率性と忠実性の評価および有効サンプルサイズ

表 3 ACE における SNIS による KL ダイバージェンスの推定及び有効サンプルサイズ

条件	ESS	ESS/n_{ac}	$\widehat{\mathrm{KL}}(g\|g')$	$\widehat{\mathrm{KL}}(g\|a')$
baseline	1	1.000	0	8.82
0-shot	38.63	0.0150	2.957	3.934
1-shot	6800	0.9982	0.002	0.002
5-shot	4081	0.5992	0.123	0.123
spec	3.311	0.0005	6.600	6.637

表 4 やさしい日本語における SNIS による KL ダイバージェンスの推定及び有効サンプルサイズ

条件	ESS	ESS/n_{ac}	$\widehat{\mathrm{KL}}(g\|g')$	$\widehat{\mathrm{KL}}(g\|a')$
baseline	187.88	0.9994	0.000	1.768
0-shot	9.65	0.0258	2.801	3.880
1-shot	1.07	0.0019	6.132	6.824
5-shot	1.73	0.0019	6.140	6.328
spec	3.84	0.0075	4.516	5.275

5.3 考察

全ての制約認識プロンプトにおいて、baseline よりも高い受理率が得られた。特に ACE ではプロン

プトによって大きな改善が見られた。以下でより詳細な結果の考察をする。

5.3.1 ACE の 1-shot/5-shot

本条件は受理率が 100%であるが、ユニークな文は数種類しか生成されていない。生成文を確認した結果、例文をコピーした出力が見られた。ESS が高いため、特定のサンプルに重みが偏って推定をしているわけではない。しかし、プロンプトの影響を強く受け多様性が他の例に比べて大きく悪化している。そのため、実際の KL ダイバージェンスは少なくとも zero-shot 以上あるとやさしい日本語の結果から推測される。

5.3.2 Zero-shot 及び CNL 仕様説明プロンプト

制限自然言語の説明を行わなかった zero-shot においても、受理率が向上した。これは LLM が ACE 及びやさしい日本語の知識を事前学習によって獲得しているといえる。制限自然言語の説明を行った spec は zero-shot に比べてより高い受理率を見せている。これは事前学習によって、知識を獲得している場合であっても、プロンプトによって文法上の制約について説明することが受理率の向上に役立つことを示唆している。

5.3.3 Few-shot プロンプト

やさしい日本語において、与えた例文の数を増やすほど受理率が高まる傾向が見られた。zero-shot と few-shot の忠実性を比較すると、1-shot でも大きく忠実性を大きく悪化させることが観察された。

5.4 制限自然言語の違い

baseline での受理率が大きく異なることが観察された。このことを制限自然言語に課されている制約や PENS 分類体系を元に考察する。やさしい日本語の判定に使用したやさにちチェッカーにおいては語彙、漢字スコアによって語彙の制約が課されており、硬さ、長さ、文法によって構文的な制約が課されている。やさしい日本語における構文的な制約は容易に満たされている。ACE の判定に使用した APE においては、APE が使用する辞書によって語彙の制約が課されている。そして ACE は談話構造表現などの形式言語に変換可能であるため、ACE は厳しい構文的制約が課せられていると考えられる。これが受理率に影響を与えていると推測される。

一方で、spec においては、ACE の方が生成確率が高い。仕様の明確な制限自然言語、すなわち Simplicity が大きい言語は、プロンプトによって大きな改善が見込める可能性がある。やさしい日本語のような明確な仕様がない言語、すなわち PENS 分類体系における Simplicity が小さい言語は、仕様の説明をするよりも few-shot のように例文を与える方が効果的な可能性がある。Precision が高く Expressiveness が小さい言語は、few-shot プロンプトの例の影響を強く受け過ぎてしまう可能性がある。

6 今後の課題

本研究では、KL ダイバージェンスを推定するために、SNIS を使用した。しかし ESS が小さく、推定結果は不安定である。そのため、より多くのサンプルで推定を行い、多様なサンプルの取得や ESS を大きくすることが必要である。実験では 1 種類の LLM を用いて実験を行った。特定の LLM モデルだけではなく、様々な LLM に対する知見を得ることが必要である。例として、リーズニングモデルとの比較が考えられる。

7 おわりに

本研究では、制限自然言語を LLM に生成させる際のプロンプトによる影響を確認した。プロンプトを与えることで制限自然言語を生成させやすくなることが観察された。その際、制限自然言語が語彙や構文及び意味にどういう制約を課すのか、その制限自然言語が PENS 分類体系においてどう分類されるかは、LLM の制限自然言語生成の性質を説明する上で有効だと考えられる。制限自然言語生成においては、baseline の受理率が低い場合がある。その場合に忠実性の評価をする方法として、自己正規化重点サンプリング (SNIS) による KL ダイバージェンス推定がある。しかし ESS が大きい場合でも、受理されたサンプル数の多様性が低く推定値よりも忠実性が悪化していると考えられる場合がある。SNIS を利用する際は ESS だけでなく、サンプルの多様性についても見ることが必要である。プロンプトによる手法は、受理率の改善に役立つものの忠実性を悪化させるため、実用性を高めるためには再学習が必須である。これまで人のために設計された制限自然言語であるが、人だけでなく LLM にとっても有用であるという観点が制限自然言語に求められるだろうと考える。

参考文献

[1] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models, 2023.

[2] 瞳谷中. ことばの意味を計算するしくみ. 講談社サイエンティフィク, 2024.

[3] Daniel Melcer, Sujan Gonugondla, Pramuditha Perera, Haifeng Qian, Wen-Hao Chiang, Yanjun Wang, Nihal Jain, Pranav Garg, Xiaofei Ma, and Anoop Deoras. Approximately aligned decoding, 2025.

[4] Tobias Kuhn. A survey and classification of controlled natural languages. **Computational Linguistics**, Vol. 40, No. 1, pp. 121–170, March 2014.

[5] Attempto Project. Attempto Project, 2025. https://attempto.ifi.uzh.ch/site/（2026-01-08 閲覧）.

[6] 宮田莉奈, 惟高日向, 山内洋輝, 柳本大輝, 梶原智之, 二宮崇, 西脇靖紘. Matcha：専門家が平易化した記事を用いたやさしい日本語パラレルコーパス. 自然言語処理, Vol. 31, No. 2, pp. 590–609, 2024.

[7] 功雄庵, 一成岩田, 篤嗣森. 「やさしい日本語」を用いた公文書の書き換え：多文化共生と日本語教育文法の接点を求めて. 人文・自然研究, Vol. 5, pp. 115–139, March 2011.

[8] 智大新井. Web ニュース記事のやさしい日本語の計量的分析. 国際日本学研究論集, Vol. 20, pp. 38–58, 2024. http://hdl.handle.net/10291/0002000796（2026-01-09 閲覧）.

[9] 出入国在留管理庁, 文化庁. 在留支援のためのやさしい日本語ガイドライン. Technical report, 出入国在留管理庁・文化庁, 8 2020. https://www.bunka.go.jp/seisaku/kokugo_nihongo/kyoiku/pdf/92484001_01.pdf（2026-01-08 閲覧）.

[10] 一成岩田, 篤嗣森, 達彦松下, 明則中島. やさにちチェッカー. https://www4414uj.sakura.ne.jp/Yasanichi1/nsindan/, 2015.（2026-01-09 閲覧）.

[11] Jens Lehmann, Preetam Gattogi, Dhananjay Bhandiwad, Sébastien Ferré, and Sahar Vahdat. Language models as controlled natural language semantic parsers for knowledge graph question answering. 2023.

[12] Jacinto A. Dávila Quintero. Controlled natural language models. In **Workshop Proceedings of the 40th International Conference on Logic Programming (ICLP-WS 2024)**, 2024.

[13] 健太郎花房, 大輝柳本, 智之梶原, 崇二宮. 大規模言語モデルによる日本語スタイル変換の性能評価. 言語処理学会 第 31 回年次大会 発表論文集, 2025.

[14] Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and Zhiyu Li. Controllable text generation for large language models: A survey, 2024.

[15] Minbeom Kim, Thibaut Thonet, Jos Rozen, Hwaran Lee, Kyomin Jung, and Marc Dymetman. Guaranteed generation from large language models, 2025.

[16] 敦志須山. ベイズ深層学習. 講談社, 2019.

A　記号の定義

文（出力）を x とする。$a(x)$ をベース LM の分布、$b(x) \in \{0,1\}$ を制約指示関数とする。

$$g(x) = \frac{a(x)b(x)}{Z}, \qquad Z = \sum_x a(x)b(x) \;=\; \mathbb{E}_{x\sim a}[b(x)],$$

$$g'(x) = \frac{a'(x)b(x)}{Z'}, \qquad Z' = \sum_x a'(x)b(x) \;=\; \mathbb{E}_{x\sim a'}[b(x)],$$

$$w(x) = \frac{a(x)}{a'(x)}, \qquad \phi(x) = \log w(x).$$

ここで Z(および Z') は正規化定数であり、b が指示関数であるため a(および a') からの棄却サンプリングにおける受理率に等しい。

B　統計手法

B.1　棄却サンプリング

棄却サンプリングは、任意の確率分布 (目的分布) と別の確率分布 (提案分布) からのサンプリングを利用して、目的分布のサンプリングをする手法である [16]。特に目的分布が $g(x) = \frac{a(x)b(x)}{Z}$ であり、提案分布が $a(x)$ であるならば、手順は通常より単純になる。言語モデル a からサンプリングして得られた x が条件 b を満たすなら、受理し、満たさないなら棄却するとすることで目的分布 g が得られる。

B.2　自己正規化重点サンプリング (Self-normalized importance sampling; SNIS)

忠実性の評価 $\mathrm{KL}(g \,\|\, g')$ は KL ダイバージェンスの定義から次のようになる。

$$\mathrm{KL}(g \,\|\, g') \;=\; \mathbb{E}_g\left[\log\left(\frac{g}{g'}\right)\right] \tag{1}$$

これは期待値であるため、g からサンプリングと単純モンテカルロ法によって推定することが可能であるが、ACE のような棄却サンプリングによって g からのサンプルを得ることが非効率な場合がある。提案分布 g' を用いて期待値を推定する手法として SNIS がある。(1) の式は次のように変形できる

$$\mathbb{E}_g\left[\log\left(\frac{g}{g'}\right)\right] = \frac{\mathbb{E}_{g'}[w\phi]}{\mathbb{E}_{g'}[w]} - \log\left(\mathbb{E}_{g'}[w]\right) \tag{2}$$

(2) の第 1 項は SNIS[16] で推定し、第 2 項は単純モンテカルロ法によって推定する。

C　実験設定の補足

C.1　プロンプト

few-shot 例文と spec 文は，表 5 から採取した。

表 5　プロンプトで使用した文の出典

区分	出典
few-shot	MATCHA，APE リポジトリのテスト文[1)]
spec	やさしい日本語ガイドラインの一部 [9]，ACE 6.7 in a Nutshell の Simple Sentences[2)]

C.2　モデル

生成モデルは `Qwen3-4B-Instruct-2507` を用いた（T=1.0, top-p=1.0, top-k=0, `max_new_tokens`=64）。

C.3　ACE における 5-shot

System prompt

```
Qwen3 must speak Attempto Controlled English. There are some examples. Qwen3, write a short sentence! there is a man. a flat mate runs. a dog is good. A man waits in a bank. A man runs and walks or sleeps.
```

User prompt

```
Qwen3, write a short sentence!
```

生成文一覧

```
1. A man waits in a bank.
2. There is a man who waits in a bank.
3. there is a man who waits in a bank.
```

C.4　やさしい日本語の合計スコア及び項目別の受理率

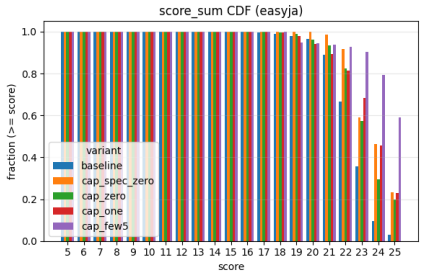

図 1　やさにちチェッカーの合計スコアの相補累積分布

表 6　やさにちチェッカーの項目ごとの受理率

条件	語彙	漢字	硬さ	長さ	文法
baseline	0.178	0.978	0.999	0.973	0.991
0-shot	0.569	0.721	1.000	0.983	0.995
1-shot	0.551	0.900	1.000	0.922	0.999
5-shot	0.848	0.926	1.000	0.994	0.999
spec	0.881	0.577	1.000	0.999	1.000

1)　https://github.com/Attempto/APE/blob/master/tests/acetexts.pl

2)　https://attempto.ifi.uzh.ch/site/docs/ace_nutshell.html