# 尤度最大化に基づく自然言語による多段推論過程の抽出への取り組み

張 辰聖子 [1]　持橋 大地 [2]　小林 一郎 [1]
[1] お茶の水女子大学大学院　[2] 統計数理研究所
{g1920524,koba}@is.ocha.ac.jp　daichi@ism.ac.jp

## 概要

ヒトが行うような自然言語による推論では，観測した事象を言語で表現することで認識を行い，その事象に関する知識を取り込んだ推論を多段に繰り返していくことにより最終的な帰結となる自然言語文を生成していると考える．本研究では，自然言語文入力に対して知識を介在させ，新たな自然言語文を生成する形で結論を導くための多段の多岐にわたる推論過程過程の中から，自然言語文を生成する際の尤度を最大化する過程を抽出する手法を提案する．実験には多段推論のデータセットである MuSiQue を利用した．結果として，提案手法はチャンスレベルを上回る性能であることを検証した．

## 1　はじめに

従来，推論は論理学において形式的なスタイルを伴うものとして研究されてきた．以前の自然言語処理分野では含意関係認識の研究 [1] がそれを代表するものとして取り上げられてきたが，大規模言語モデルの出現により推論自体を自然言語で行うことが可能になった [2, 3, 4]．言語による推論とは，観察対象の言語による認識からそれに対する因果的な帰結を説明するという形で行われ，それは前提となる情報から，結論を言語で表現するという自然言語文生成とみなすことができる．この際，一般的に前件の観察対象の情報に加えて，それに関する知識を踏まえて帰結となる自然言語文を生成することが想定される．知識は断片的に存在しているのではなく，言語の使用と結びついており，ヒトによる言語を用いた観測事象の認識は，認識に利用可能な知識を踏まえてなされており，推論は認識状態に適用可能な因果的知識によって表現されると考える．これを可能性の高い過程の元で繰り返し適用することで，ヒトは言語による思考を実現していると考える．このことから，本研究では前提から結論を導くにあたって想定される知識を，言語モデルの生成確率を用いて選択することで，前提から結論への適切な推論過程を生成する手法を提案する．

実験を通じて提案手法を検証するために，多段推論のためのデータセットである MuSiQue [5] の質問をもとに言語モデルを用いて知識を複数生成し，結論への生成確率で重み付けを行い，次の段階の知識を生成するという particle filter の考え [6] を用いて推論過程の抽出を行う．

## 2　関連研究

大規模言語モデルを用いて推論を行う際に，プロンプトを用いて推論過程を与える手法が近年活発に研究されている．大規模言語モデルの強力な文脈内学習能力 [7] を考慮し，推論過程を生成させるために，一連の中間推論過程を Few-Shot で与える Chain-of-Thought prompting [8] や，“Let’s think step by step!”と与えるだけで，Zero-Shot で推論過程を生成する手法 [9] などがある．このように，推論においてプロンプトの付与の仕方によって推論過程の精度を向上させることで，回答を良いものにする研究が多くある．この際，生成された推論過程を，回答から帰納的に評価し学習を行う手法として，生成された推論過程と質問と回答との類似度をスコアとして学習を行う手法 [10] や，生成された推論過程に対して，正しい答えを導くものであれば，比較的妥当な推論過程であると判断して，データセットとしてファインチューニングを行う手法 [11]，生成された知識を加えることによる回答の変化から PPO[12] に基づいて報酬づけを行なって学習させる手法 [13] がある．これらに対し，我々は回答の生成確率から，帰納的に推論過程を評価する．

本研究では，ヒトが自然言語で推論する認知活動に着目し，LLM の中に存在する因果関係を示す潜在

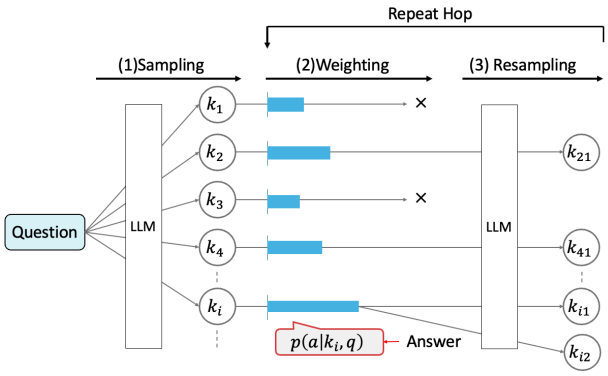


**図 1**　推論過程の生成の概要.particle filter の考え [6] を用いて，Sampling，Weighting，Resampling を繰り返し行い推論過程を生成する．

的な知識構造を自然言語文を生成する過程を自然言語推論過程と見做し，潜在的知識構造を生成確率が高い自然言語文として部分的に表出することで適切な推論過程を抽出する手法を提案する．

# 3　推論過程の生成

図 1 に研究の概要を示す．質疑から回答を導く推論タスクを展開し推論過程を生成する．この際，推論過程は particle filter と同じ推論過程となる．以下，particle filter のパーティクルを生成される自然言語文と考えて説明を行う．

## 3.1　データセット

本研究では，MuSiQue [5] を利用する．コンテキストの情報から生成されたシングルホップの質問を組み合わせることで，マルチホップの質問を生成しているデータセットである．このデータセットの特徴として，前段階の hop の回答が次の段階の hop の質問に含まれるという推論間の依存性と，コンテキストの情報を用いることで多段の質問に答えるという，コンテキストへの依存性があげられる．

## 3.2　研究概要

**(1) 予測**

言語モデルに例と質問を与える few-shot のプロンプトを入力することで，質問に沿った知識を生成させる．ここで与えるプロンプトは，言語モデルから有用な知識を生成する手法 [15] をもとにしている．知識を生成する際に使用する質問と知識のペアの例は，train データの中からランダム抽出している．プロンプトの一部を表 3 に示す．このプロンプトは，知識を生成させる指示，データセットに沿った適切な入力と出力の例，および与える質問を挿入するプレースホルダで構成される．新しい質問に対する知識を生成する場合，プロンプトのプレースホルダに代入し，言語モデルに与えることで，知識を生成する（サンプリング）．コンテキストを与える場合は，冒頭に挿入する．これは particle filter における予測部分に対応し，初期状態（question）からシステムモデル（LLM）にしたがって予測サンプル（知識）を生成する動作に準ずる．

**(2) 尤度計算**

質問 $q$ に対して，$N$ 回の hop を通じて回答 $a$ が生成される確率は，$n$ 回目の hop において質問に関連して生成される無数の知識を $k_i$ として，式 (1) で表される．

$$
\begin{aligned}
p(a|q) &= \sum_{n=1}^{N} \sum_{k_i} p(a, k_i|q) \\
&= \sum_{n=1}^{N} \sum_{k_i} p(a|k_i, q) p(k_i|q)
\end{aligned} \tag{1}
$$

式 (1) から，質問から回答を生成する確率は，質問から知識が生成される確率と，質問と知識から回答が生成される確率の積となることがわかる．そこで，質問と知識から回答が生成される確率が高い推論過程であれば，回答を導くのに適した推論過程であるとして $p(a|k_i, q)$ の重み付けを行う．この確率を言語モデルの生成確率で算出する．具体的には，言語モデルに文を与え，各トークンに対する予測確率を取得し，質問と知識が与えられた時の回答に対応する条件付き確率を算出している．particle filter における尤度計算の部分に対応し，観測値（回答 $a$）と予測値（知識 $k$）との差分に基づき，各サンプルの尤度（生成確率）を計算する動作に準ずる．

**(3) リサンプリング**

尤度に比例して次の段階で生成する知識の数を決定し，前段階の知識から次の段階の知識を生成する．言語モデルに表 3 のプロンプトに前段階までの知識を追加したプロンプトを与えることで，質問と前段階の知識に沿った知識を生成させる．particle filter におけるリサンプリング部分に対応し，重み（尤度）に比例した個数のサンプル (知識) を生成する動作に準ずる．

(2) 尤度計算，(3) リサンプリングを繰り返すことで，質問から，回答を導く多段推論の推論過程を生成する．

**表 1**　推論過程に対する BERT-Score と BLEU.

| 推論過程 | 生成手法 | BERT-Score Precision 元文 | BERT-Score Precision 平叙文 | BERT-Score Recall 元文 | BERT-Score Recall 平叙文 | BERT-Score $F_1$ 元文 | BERT-Score $F_1$ 平叙文 | BLEU 元文 | BLEU 平叙文 |
|---|---|---|---|---|---|---|---|---|---|
| hop-1 | CoT | 0.016 | 0.104 | 0.002 | 0.177 | 0.009 | 0.140 | 0.017 | 0.025 |
| | +context | 0.061 | 0.141 | 0.071 | 0.233 | 0.066 | 0.187 | 0.014 | 0.021 |
| | 5-shot | 0.118 | 0.240 | 0.061 | 0.288 | 0.089 | 0.263 | 0.016 | 0.056 |
| | +context | 0.162 | 0.280 | 0.089 | 0.311 | 0.125 | 0.294 | 0.016 | 0.054 |
| | ours | 0.154 | 0.267 | 0.080 | 0.301 | 0.116 | 0.283 | 0.017 | 0.057 |
| | +context | 0.168 | 0.287 | 0.103 | 0.338 | 0.134 | 0.311 | 0.020 | 0.063 |
| hop-2 | CoT | −0.001 | 0.058 | −0.042 | 0.105 | −0.021 | 0.082 | 0.008 | 0.010 |
| | +context | 0.020 | 0.076 | −0.002 | 0.146 | 0.010 | 0.111 | 0.011 | 0.012 |
| | 5-shot | 0.051 | 0.150 | −0.009 | 0.197 | 0.021 | 0.173 | 0.007 | 0.018 |
| | +context | 0.124 | 0.238 | 0.061 | 0.286 | 0.092 | 0.269 | 0.013 | 0.057 |
| | ours | 0.094 | 0.191 | 0.015 | 0.214 | 0.054 | 0.202 | 0.012 | 0.025 |
| | +context | 0.124 | 0.232 | 0.054 | 0.275 | 0.088 | 0.253 | 0.015 | 0.052 |
| hop-3 | CoT | 0.056 | 0.085 | 0.047 | 0.137 | 0.052 | 0.111 | 0.013 | 0.014 |
| | +context | 0.149 | 0.112 | 0.135 | 0.181 | 0.142 | 0.146 | 0.015 | 0.023 |
| | 5-shot | 0.149 | 0.215 | 0.135 | 0.270 | 0.142 | 0.242 | 0.012 | 0.027 |
| | +context | 0.192 | 0.270 | 0.135 | 0.332 | 0.185 | 0.301 | 0.016 | 0.058 |
| | ours | 0.208 | 0.292 | 0.164 | 0.312 | 0.186 | 0.301 | 0.018 | 0.046 |
| | +context | 0.287 | 0.392 | 0.252 | 0.431 | 0.269 | 0.410 | 0.044 | 0.111 |
| 平均 | CoT | 0.024 | 0.082 | 0.002 | 0.139 | 0.013 | 0.111 | 0.012 | 0.016 |
| | +context | 0.077 | 0.110 | 0.068 | 0.186 | 0.073 | 0.148 | 0.013 | 0.019 |
| | 5-shot | 0.106 | 0.202 | 0.062 | 0.252 | 0.084 | 0.226 | 0.012 | 0.034 |
| | +context | 0.159 | 0.263 | 0.095 | 0.310 | 0.134 | 0.288 | 0.015 | 0.056 |
| | ours | 0.152 | 0.250 | 0.076 | 0.268 | 0.119 | 0.262 | 0.016 | 0.042 |
| | +context | **0.193** | **0.304** | **0.136** | **0.348** | **0.164** | **0.325** | **0.026** | **0.076** |

**表 2**　推論過程に対する G-Eval [14] の結果.

| | consistency | coherence | relevance |
|---|---|---|---|
| CoT | 2.46 | 2.78 | 2.81 |
| +context | 3.21 | 3.41 | 3.59 |
| 5-shot | 2.16 | 1.58 | 2.13 |
| +context | 2.65 | 3.07 | 3.11 |
| ours | 2.50 | 1.91 | 2.51 |
| +context | **3.29** | **3.66** | **3.67** |
| 正解 | 4.76 | 4.22 | 4.69 |

# 4　実験

## 4.1　実験設定

3.1 節にて紹介したマルチホップデータセットである MuSiQu から 3hop で直線的なグラフ構造を持つ 567 個の評価データを対象にする．また，推論過程が質疑として与えられているのに対し生成文は平叙文なので，GPT-4 を用いて質疑として与えられる推論過程を平叙文に直したものに対する評価も行なった．生成される知識の数を 10 とし，推論過程の生成を行う．評価方法として，CoT として「Let's think step by step」と与えて推論過程を出力した場合と，5-shot で生成例を与えて推論過程を出力した場合と提案手法を比較する．加えて，コンテキストを含める場合と含めない場合を比較する．評価指標は，BERT-Score[16] と BLEU[17] の文の類似度を測る 2 つに加え，GPT-4 を用いた評価方法である G-Eval[14] を用いる．G-Eval では推論過程について coherence, consistency, relevance を評価させる．ここでは，生成データに対する評価だけでなく，正解データに対しての評価も行わせることで，評価指標の正当性を示す．本研究で用いるモデルは Hugging Face[1] の自然言語ライブラリ Transformers に基づく，Meta 社公開の事前学習済み Llama3.1 モデル[2] を使用した．

## 4.2　実験結果

生成された知識の例を表 3 に示す．このように，質問に対して段階的に知識を生成し，推論過程に対して評価をおこなう．表 1 に BERT-Score と BLEU の結果を，表 2 に G-Eval による結果を示す．図 2 に

1）　https://huggingface.co
2）　https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

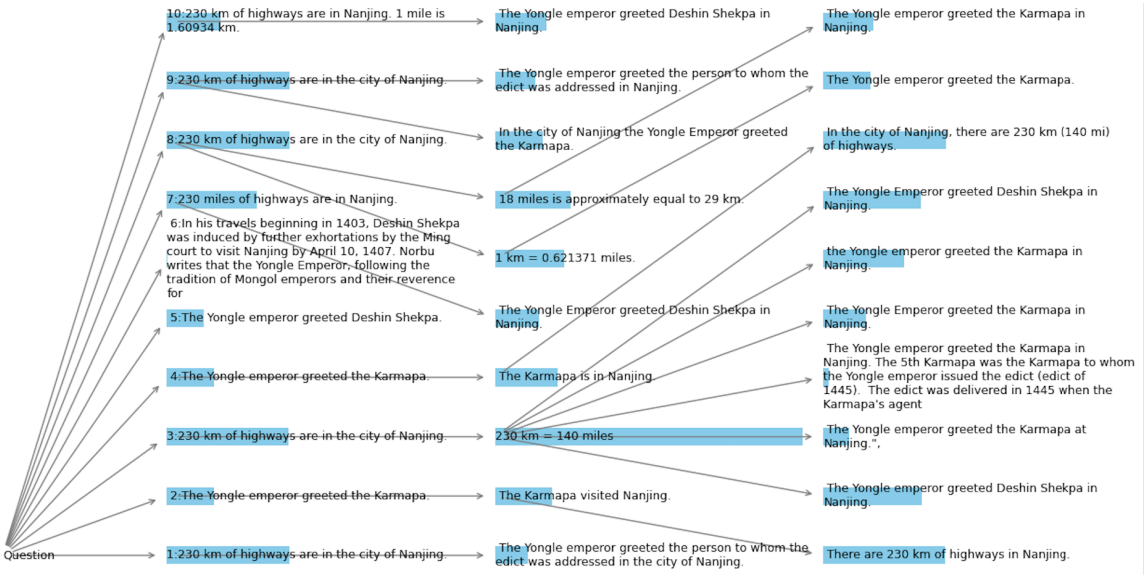


**図 2**　コンテキストを与えた場合の提案手法で生成された知識と，それらに対する尤度を棒グラフで表す．

表 3 中の質問に対応する提案手法にコンテキストを加えた場合での，生成された知識と，それらに対する尤度を可視化させたもの示す．BERT-Score, BLEU スコアどちらの評価指標においても，コンテキストを含めた提案手法が最も良い結果となった．

## 5　考察

提案手法により推論過程を選択することで，CoT や 5-shot で生成したものに比べ回答に沿った推論過程を抽出することができた．表 1，表 2 から，コンテキストを加えた場合と加えていない場合のどちらをみても，類似度をはかる指標である BERT-score, BLEU に加え，LLM によって評価されたほとんどの結果で，CoT や 5-shot での生成よりも提案手法によって生成された推論過程の方が良い結果となった．このことから，回答への生成確率を解析しながら推論過程を生成することの有効性が示せる．

今回使用したマルチホップデータセットの特徴として，hop 同士の依存性に加え，コンテキストへの依存性が挙げられる．そこで，コンテキストがある場合とない場合とでどれほど精度に差が生まれるか実験を行った．その結果，表 1，表 2 から提案手法の場合でも CoT や 5-shot の場合でもコンテキストを加えることで全ての結果が向上した．表 3 において，コンテキストを与えていない状態の提案手法による出力とコンテキストを与えた提案手法による出力に注目すると，どちらとも正解の多段推論のようなスッテプを踏んで推論を行なっているが，コンテキストを与えていない状態の提案手法による出力の場合，コンテキストの情報がなく LLM 内にある情報に依存するため，間違った内容になっている．これらのことから，コンテキストへの依存度を示すことができた．

多段推論としての知識を生成する際に起こりうる問題として，質問に対して，多段となる知識ではなく回答を導く直接的な知識を生成してしまうというものがある．実際に，表 3 においてコンテキストを加えた 5-shot で生成された出力結果を見ると，hop-1 や hop-2 で hop-3 のように直接的に回答につながる知識を生成してしまっている．

## 6　おわりに

本研究では，前提から生成される多岐にわたる推論過程の中から，結論を導くために用いられたものを大規模言語モデルの尤度によって予測する方法を提案し，開発した．マルチホップデータセットの質問に対し，few-shot で複数の知識を生成し，言語モデルの生成確率を用いて回答への尤度から重み付けを行い，それを元に次の段階の知識を生成することで回答に沿った推論過程を抽出する．この手法により，ランダムで推論過程を生成したものに比べ，BERT-score，BLEU スコアによる類似度，および LLM を用いた評価の双方で精度を向上させることができた．今後は，問題点としてあげた直接的な知識の生成を抑制するため，パラメータの調整や，プロンプトに工夫を与えてみたい．また，今回は機械的評価しか行っていないため，人間による評価でどのような違いが見られるかも確認したい．

## 謝辞

本研究は JSPS 科研費 JP23K28143 の助成を受けたものです.

## 参考文献

[1] Aarthi Paramasivam and S. Jaya Nirmala. A survey on textual entailment based question answering. **Journal of King Saud University - Computer and Information Sciences**, Vol. 34, No. 10, Part B, pp. 9644–9653, 2022.

[2] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. **arXiv preprint arXiv:2212.10403**, 2022.

[3] Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. Natural language reasoning, a survey, 2023.

[4] Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. Reasoning with language model prompting: A survey. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 5368–5393, Toronto, Canada, July 2023. Association for Computational Linguistics.

[5] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. **Transactions of the Association for Computational Linguistics**, 2022.

[6] Pierre Del Moral. Nonlinear filtering : Interacting particle resolution. **Comptes Rendus De L Academie Des Sciences Serie I-mathematique**, Vol. 325, pp. 653–658, 1997.

[7] Yihan Wang, Si Si, Daliang Li, Michal Lukasik, Felix Yu, Cho-Jui Hsieh, Inderjit S Dhillon, and Sanjiv Kumar. Preserving in-context learning ability in large language model fine-tuning, 2023.

[8] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[9] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023.

[10] Veronica Latcinnik and Jonathan Berant. Explaining question answering models through text generation, 2020.

[11] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning, 2022.

[12] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

[13] Jiacheng Liu, Ramakanth Pasunuru, Hannaneh Hajishirzi, Yejin Choi, and Asli Celikyilmaz. Crystal: Introspective reasoners reinforced with self-feedback, 2023.

[14] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023.

[15] Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. Generated knowledge prompting for commonsense reasoning, 2022.

[16] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In **International Conference on Learning Representations**, 2020.

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

# 付録 A

**表 3** 質問に対して生成された知識の例と正解．各 hop の平叙文に対する出力の BERT-Score を右に示す．

元文

**Q:** How many miles of highways are in the city where the Yongle emperor greeted the person to whom the edict was addressed?
**A:** 140 mi

Prompt

Generate some knowledge about the input. Examples:

Input: Who is the father of the Eleanor of the country where Eveline Adelheid von Maydell died?
Knowledge: Eveline Adelheid von Maydell died at Sintra.
...
Input: {**question**}
Knowledge:

| hop | 生成手法 | 推論過程 | BERT-Score(F1) |
|---|---|---|---|
| 1 | 正解 | **Q:** Who was the edict addressed to?<br>**A:** the Karmapa<br>→**平叙文:** The edict was addressed to the Karmapa. | |
| | 5-shot | There is a Yongle Emperor related Chinese map of the world. | 0.068 |
| | +context | 230 km of highways are in the city of Nanjing, and the Yongle emperor greeted the Karmapa. | 0.276 |
| | ours | The person to whom the edict was addressed was Vân Tng. | 0.272 |
| | +context | The edict was addressed to the Karmapa. | 1.000 |
| 2 | 正解 | **Q:** Where did the Yongle Emperor greet the the Karmapa ?<br>**A:** Nanjing<br>→**平叙文:** The Yongle Emperor greeted the Karmapa in Nanjing. | |
| | 5-shot | The Yongle Emperor sent an edict to Zheng He. | 0.410 |
| | +context | 230 km is equal to 140 miles. | −0.084 |
| | ours | Vân Tng is located in Beijing. | −0.240 |
| | +context | The Yongle emperor greeted the Karmapa in Nanjing. | 0.975 |
| 3 | 正解 | **Q:** How many miles of highways are in Nanjing ?<br>**A:** 140 mi<br>→**平叙文:** Nanjing has 140 miles of highways. | |
| | 5-shot | Zheng He arrived in Malacca, a city of Malaysia. | 0.202 |
| | +context | 140 miles of highways are in the city of Nanjing. | 0.494 |
| | ours | Beijing has highway mileage of 4,179 miles. | 0.388 |
| | +context | In the city of Nanjing, there are 230 km (140 mi) of highways. | 0.486 |