自然言語による帰納的推論過程の抽出

張 辰聖子　(指導教員：小林 一郎)

1　はじめに

人間が行っている推論は論理的な含意関係だけではなく，日常的な知識を背景にした常識推論が大きな役割を果たしている．本研究はこの際の知識を，複雑な内容を表現できる自然言語文自体ととらえ，「から」「ため」のような手がかり表現を利用することで，前提と帰結がともに文となる自然言語推論をコーパスから深層学習モデルとして直接学習する．加えて，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．

2　実験手法

2.1　手がかり表現に基づく非論理的な言語推論の学習

日本語 Wikipedia コーパスから,「理由」を表現する手がかり表現を元に根拠と結論がペアになった文を抽出する．根拠部分を入力，結論部分を出力とし，文を生成する深層学習モデルである T5[3] に学習させる．これにより根拠に相当する文から結論を示す文を生成することを通じて，自然言語推論を実現する．

2.2　言語モデルの生成確率から選択された推論過程による学習

質疑を推論タスクとみなし，推論過程として知識を与える．質問 q に対して，回答 a が生成される確率は，質問に関連する無数の知識を k として，式 (1) で表される．

$$\begin{aligned} p(a|q) &= \sum_k p(a,k|q) \\ &= \sum_k p(a|k,q)p(k|q) \end{aligned} \tag{1}$$

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．このことから，データ生成として「Yahoo! 知恵袋データ」の質疑に対し，言語モデルを用いて質問に対する理由を生成し，生成確率を元にデータセットに追加していく．次にモデルの学習として，質問を入力として知識を生成するモデルと，質問と生成された知識を入力として回答を生成するモデルを，文を生成する深層学習モデルである GPT2[2] に学習させることで構築する．これにより，回答の精度が上がるかを確認し，生成確率による推論過程の正当性を示すのが本提案の目標である．

2.3　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

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，推論過程はパーティクルフィルタと同じ推論過程となる．以下，パーティクルフィルタ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．

まず予測を行い，質問に対して言語モデルを用いて

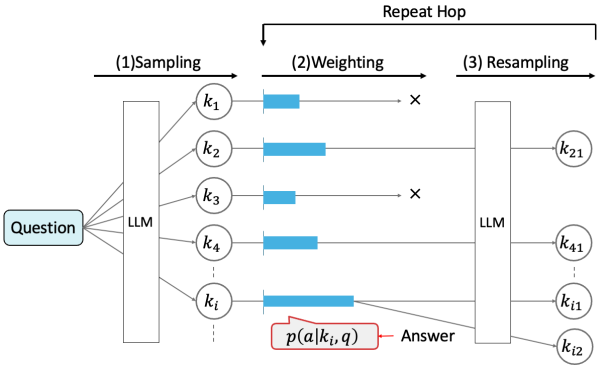

図 1: 推論過程の生成の概要．

質問に対する知識をランダムに複数個生成する．次に尤度計算を行い，予測で生成された各知識に対して，言語モデルの生成確率を用いて回答文の尤度を計算する．最後にリサンプリングとして，尤度に比例した数に応じて前段階の知識から次の段階の知識を生成する．これを繰り返すことで，大規模言語モデルが有する潜在的な知識を因果関係の元に抽出する，質問から最終的な回答に至る推論過程を示す自然言語文の生成を行う．

3　実験

3.1　手がかり表現に基づく非論理的な言語推論の学習の実験

実験設定　この実験では，2.1 節の手法で収集した 1,572,956 件の根拠と結果のペア文をデータとして用いる．実験の際は，このデータを訓練：開発：評価 =0.95:0.025:0.025 として評価を行う．評価は評価データからランダムで 100 文を抽出し手動で行い，それに伴い BERT-Score[4] によって，生成文と正解文との意味的類似度を検証した．本研究で用いる T5 モデルは Hugging Face の自然言語ライブラリ Transformers に基づく，Isao Sonobe 公開の事前学習済み日本語 T5 モデル[1]を使用した．

実験結果・考察　ここでは要旨のため，生成データセットにおけるモデルの性能に対する人手で行った評価結果のうち，全ての評価者 4 名の平均をグラフにしたもののみを図 2 に示す．

実験の結果，前提を与えることで，文生成の形で推論を行うことができた．出力が妥当な推論である，もしくは妥当な推論であるが文法的な間違いがあると評価された割合は 65.8 %となった．

3.2　言語モデルの生成確率から選択された推論過程による学習の実験

実験設定　実験にあたっては，予備実験として小さなデータから始めるため，「Yahoo!知恵袋データ」から「> 健康，美容とファッション > 健康，病気，病院」カテゴリの投稿のみをデータとして用いる．また，トークン数は 128 トークン以下という制約を設け，訓練データを 1000 ペア，評価データを 500 ペアとして訓練，評

[1]https://huggingface.co/sonoisa/t5-base-japanese

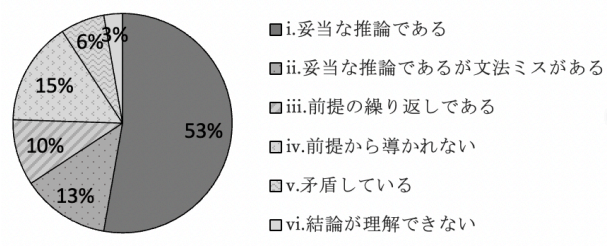

図 2: 評価データからランダムに抽出した 100 文に対する人手評価の結果 (平均)

価を行う．また，訓練で用いる推論過程の知識として 2.2 節で収集した知識を，生成数を 10 とし，動作を繰り返す回数を 3 とする．これにより,「質問，回答，知識」のデータセットを 3000 セット用意し，これを用いて学習を行う．

評価は，BERT-Score[4] と BLEU[1] の 2 つの評価指標を用いて行う．本研究で用いる GPT2 モデルは Hugging Face の自然言語ライブラリ Transformers に基づく，rinna 社公開の事前学習済み日本語 GPT2 モデル[2]を使用した．

以下の 4 つのモデルに対して評価を行った．知識を用いずに，質問に対して直接 GPT2 で回答生成を行ったものを Direct GPT2 とする．言語モデルによって Few-Shot で知識を生成し，質問に加えた形に対して GPT2 で回答生成を行ったものを知識あり GPT2 とする．また，提案手法で生成したデータセットでファインチューニングを一度行ったものを iter1，二度行ったものを iter2 とする．

実験結果・考察　表 1 に実験結果を示す．BERT-Score, BLEU スコアどちらの評価指標においても，生成したデータセットによってファインチューニングを行った iter1 の結果が最も高い結果となった．提案手法により生成したデータセットによってファインチューニングを行うことで，精度の高い知識と回答の生成ができるようになり，知識を与えることによる回答への影響も確認できた．

3.3 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の実験

実験設定　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．本研究で用いる Llama3.1 モデルは Hugging Face 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[3]を使用した．

[2]https://huggingface.co/rinna/japanese-gpt-1b

[3]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

表 1: 回答生成の実験結果.

モデル	BERT-Score(F_1)	BLEU
Direct GPT2	0.613	0.081
知識あり GPT2	0.613	0.081
iter1	**0.619**	**0.085**
iter2	0.604	0.082

表 2: 推論過程に対する BERT-Score(F1) と BLEU.

	BERT-Score(F_1)		BLEU	
生成手法	元文	平叙文	元文	平叙文
CoT	0.013	0.111	0.012	0.016
+context	0.073	0.148	0.013	0.019
5-shot	0.084	0.226	0.012	0.034
+context	0.134	0.288	0.015	0.056
ours	0.119	0.262	0.016	0.042
+context	**0.164**	**0.325**	**0.026**	**0.076**

実験結果・考察　表 2 に BERT-Score と BLEU の実験結果を示す．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．提案手法により推論過程を選択することで，ランダム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．また，コンテキストへの依存性や，知識生成時に起こりうる問題についても調べることができた．

4 まとめ

本研究では，自然言語文そのままの形で推論を行う手法の開発を目的に，自然言語による推論を自然言語文生成として表現する研究に取り組んだ．日本語 Wikipedia コーパスから，理由節を手がかり表現として根拠と結論を抽出した．抽出した因果関係のデータを用いて，根拠部分を入力，結論部分を出力として深層学習を行うことで，推論を行う形の文生成が可能なことがわかった．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ことで，より適切な推論過程を生成することができた．今後は，プロンプトやパラメータに工夫を与えることで，生成文の精度向上を試みたり，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．また，さらに複数のホップであったり，枝分かれが起こるような推論過程ではどのような結果になるか，検証を行いたい．

参考文献

[1] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: Bleu: a Method for Automatic Evaluation of Machine Translation, in Isabelle, P., Charniak, E. and Lin, D. eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA (2002), Association for Computational Linguistics.

[2] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I.: Language Models are Unsupervised Multitask Learners (2019).

[3] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P. J.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, *Journal of Machine Learning Research*, Vol. 21, No. 140, pp. 1–67 (2020).

[4] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. and Artzi, Y.: BERTScore: Evaluating Text Generation with BERT, in *International Conference on Learning Representations* (2020).