大規模言語モデルにおける文生成方向に関する依存性の検証

谷口 大輔[1]　脇本 宏平[2]　丹羽 彩奈[1]　岡崎 直観[1]
[1] 東京工業大学　[2] 株式会社サイバーエージェント
{daisuke.taniguchi@nlp.,ayana.niwa@nlp.,okazaki@}c.titech.ac.jp
wakimoto_kohei@cyberagent.co.jp

概要

大規模言語モデルは様々な自然言語処理タスクで使用され，高い性能を示してきた．ところが，事柄が記述される順序は言語によって偏りがあるため，通常の左から右に（順方向に）単語を予測する言語モデルではなく，右から左に（逆方向に）単語を予測するように学習した言語モデルの方が解きやすいタスクが存在するかもしれない．本研究では，言語モデルがタスクを解くときに単語の生成方向が与える影響を検証するため，順方向あるいは逆方向に学習した二種類の大規模言語モデルを常識推論タスクに適用し，その性能の比較・分析を行う．そして，これまで十分に利用が検討されてこなかった逆方向言語モデルの可能性を報告する．

1　はじめに

近年，GPT-3[1] をはじめとする大規模言語モデルは，様々な自然言語処理タスクで高い性能を示している．言語モデルは，単語列に対して確率を計算するものであり，ある単語列が与えられた時，その次に続く単語を一つずつ予測する処理を繰り返すことで，単語列を逐次的に生成できる．x を入力文，s_i を入力文中のトークン，n を入力文の長さとした時，言語モデルの学習は，コーパス上で

$$p(x) = \prod_{i=1}^{n} p(s_i|s_{<i}) \tag{1}$$

という式を最大化することで行われる．

ここで，言語において事柄が記述される順序には偏りがあるため，言語モデルの学習コーパスにはその偏りが内在しているはずだという点に着目する．例えば日本語では，理由を説明した後に結果や結論を記述する傾向があるため [2]，日本語のコーパスはその傾向を反映し，理由などの前提が先，結論が後に記述される文を多く含むと考えられる．一方で，結論を先，理由を後に記述する文は相対的には少ないと考えられる．自己回帰型の大規模言語モデルの学習は，過去のトークン列から現在のトークンを予測するという部分に時刻に関する非対称性があるため，単語の並び順に関する順序の偏りの影響を受けていると考えられる．そこで，本研究ではテキストを順方向に入力して学習した通常の言語モデルではなく，逆方向に入力して学習した言語モデルの方が解きやすいタスクが存在すると考える．

日本語での具体例として，何らかの結論を表す文が与えられ，それに対して尤もらしい理由を説明する文を生成する常識推論のタスクを考える．前述の通り，日本語では理由が先，結論が後で記述される文が多いため，言語モデルの単語の生成方向を考えると，結論で条件付けして理由を生成するタスクは，順方向よりも逆方向の言語モデルの方が解きやすいと予想される．これまでにも，逆方向の言語モデルで文を生成する研究 [3, 4] は存在していたが，同一タスクを順方向・逆方向の言語モデルで解き，その性能差を単語の並び順の偏りに位置付けて検証した研究はない．

本研究では，言語モデルにおける単語の生成方向がタスク処理に与える影響を検証するため，日本語コーパスで学習した GPT-2 [5] medium 相当の順方向および逆方向の言語モデルを利用する．常識推論のデータセットである Choice of Plausible Alternatives (COPA) [6] を人手で日本語に翻訳したもの[1)]を評価データとし，zero-shot による生成的な常識推論タスクに両モデルを適用し，その性能の比較・分析を行う．その実験結果から，これまで利用が十分に検討されてこなかった逆方向言語モデルの方が，通常の順方向言語モデルよりも有利に処理できる文生成タスクが存在することを報告する[2)]．

1）https://github.com/nlp-titech/copa-japanese
2）BERT 等のマスク言語モデルは双方向の言語モデルだが，文生成タスクには適していないため，本研究では扱わない．

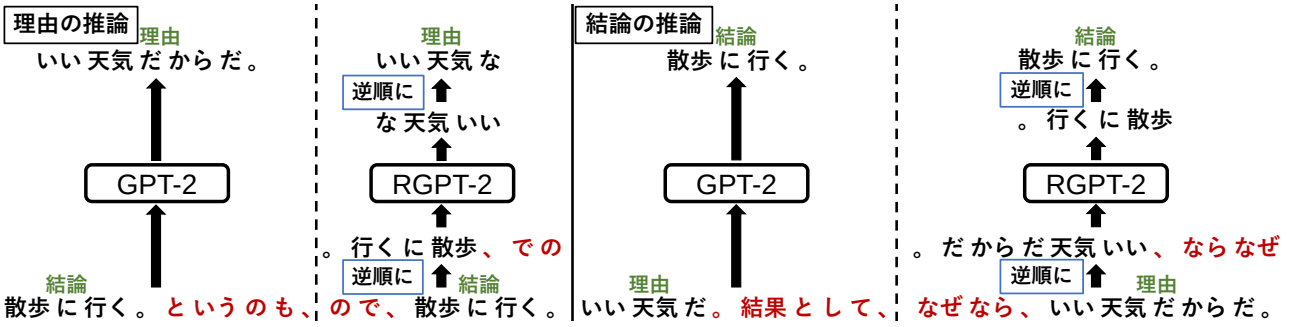

図 1　順方向（GPT-2）および逆方向（RGPT-2）の言語モデルの使用例・タスク例

2　検証方法

本研究では，単語の生成の順番が順方向および逆方向の大規模言語モデルを学習する．両モデルとも，MeCab と WordPiece[7] による東北大 BERT-v2[3] のトークナイザを利用し，GPT-2 medium 相当の日本語言語モデルを学習した．

2.1　逆方向言語モデル

本研究で用いる逆方向言語モデルを説明する．通常の順方向言語モデルの学習は，式 1 の最大化によって行われるが，逆方向言語モデルの場合は以下の式 2 の最大化を行うこととなる．

$$p(x) = \prod_{i=n}^{1} p(s_i | s_{>i}) \tag{2}$$

つまり，未来のトークン列から現在のトークンを予測する学習が行われ，結果として，文末から文頭の向きに単語列を生成するモデルを構築できる．

以降，本研究で用いる順方向言語モデルを GPT-2，逆方向言語モデルを Reversed GPT-2 (RGPT-2) と称する．RGPT-2 の実際の学習では，入力文を順方向言語モデルと同じトークナイザによって処理した後，トークン列を逆順に並び替えてモデルへの入力とし，通常の言語モデルの学習を行うことにより，実質的に式 2 の最大化を行う．

2.2　生成的な常識推論タスク

GPT-2 と RGPT-2 の性能比較を行うため，生成的な常識推論タスクを考える．具体的には，結論の文が与えられ，それに対する理由を生成するタスクと，それとは逆に，理由の文が与えられ，それに対する結論を生成するタスクの二つである．これら二つのタスクは，入力と出力が入れ替わったタスク対であるため，言語モデルの文生成方向に関する依存性により，各タスクでモデルの性能に差が生じると予想される．GPT-2 と RGPT-2 がこれらのタスクを解く例を図 1 に示す．

結論から理由を推論するタスク　理由などが先で，結論が後で記述されやすいという，日本語における説明の順序の偏りにより，RGPT-2 の方が GPT-2 よりも解きやすいと考えられる．そこで，与えられた結論に対して，言語モデルによる理由の生成を zero-shot にて行う．各モデルが出力した理由が，結論に対する理由付けとして適切かどうかを評価する．

理由から結論を推論するタスク　GPT-2 の方が RGPT-2 よりも解きやすいと考えられる．与えられた理由に対して，言語モデルによる結論の生成を zero-shot にて行う．各モデルが出力した結論が，理由から導かれる結論として適切かどうかを評価する．

2.3　プロンプティング

本研究では，理由から結論，および結論から理由を生成するタスクを，プロンプトによる zero-shot 生成タスクとして解く．図 1 に GPT-2 および RGPT-2 によるタスクの解き方を示す．なお，図中の赤字は，接続表現を表す．

GPT-2　与えられる結論（または理由）に続けて，接続表現を連結したものをプロンプトとし，それに続く単語列として理由（または結論）を出力する．

RGPT-2　与えられる結論（または理由）の直前に，接続表現を連結したものをプロンプトとし，これをトークン単位で逆順にした上で，言語モデルに入力する．その続きをモデルに生成させ，その出力をトークン単位で逆順にしたものが，理由（または

3）https://huggingface.co/cl-tohoku/bert-base-japanese-v2

結論）の出力となる．

本研究では，GPT-2 と RGPT-2 の常識推論における能力を公平に比較したいので，用いるプロンプトによる不公平をできるだけ排除する必要がある．理由を生成するタスクを考えると，GPT-2 では「（結論）**なぜなら，**（理由）」のようなプロンプト，RGPT-2 では「（理由）**ので，**（結論）」のようなプロンプト[4]を用いることになる．GPT-2 と RGPT-2 で同じプロンプトを用いることが理想ではあるが，（日本語でそのような言い方がされないため）現実には不可能で，実験に用いるプロンプトは両者で異なるものにならざるを得ない．

そこで，順方向と逆方向の言語モデルの比較において，常識推論に優れたプロンプトの存在も利点・欠点に含まれると考え，両者の言語モデルにとってできるだけよいプロンプトを選び，実験を行うこととする．

3 実験

3.1 実験設定

大規模言語モデル GPT-2，RGPT-2 の学習には，rinna 株式会社[5]が提供しているコードベース[6] を使用した．学習コーパスはコードベースに従い，日本語の CC-100 と，Wikipedia のダンプ[7]を使用した．これらは自然な日本語文を多く含み，日本語において事柄が記述される順序の偏りを内包していると期待される．このコーパス上で両モデルをそれぞれ，バッチサイズ 3 で 4,680,000 ステップ学習した．検証用データにおけるパープレキシティはそれぞれ，9.80，9.79 で，同程度のパープレキシティと言える．

データセット 生成的な常識推論のデータセットとして，Choice of Plausible Alternatives (COPA) [6] を人手で日本語に翻訳したものを使用する．COPA は本来，原因（または結果）を表す前提文に対する結果（または原因）としてふさわしい文を，二つの選択肢から選ぶ問題であるが，本実験では，そのうちの正解文と前提文のみを抽出し，理由（原因）と結論（結果）の文ペアとして用いる．また，GPT-2 と RGPT-2 の性能を極力公平に評価するため，「彼（ら）」，「彼女（ら）」，「それ（ら）」，「そこ」といった代名詞が理由・結論のいずれにも含まれないペアのみを評価対象とした[8]．開発データは 186 件，評価データは 182 件である．

接続表現 理由・結論のプロンプティングに利用する接続表現の候補を，言語モデルの学習コーパスの中から手動で収集した．GPT-2 による理由の推論と RGPT-2 による結論の推論，および GPT-2 による結論の推論と RGPT-2 による理由の推論で同じ接続表現の候補を利用することとし，各 8 件ずつ候補を用意した．それらの接続表現が学習コーパス内で出現した回数を付録 A に示す．実験では，開発データの全事例における zero-shot 生成の BLEU スコアが高くなるものを採用し，両モデルの性能を比較する．

自動評価 自動評価指標として，BLEU スコアとパープレキシティを使用する．BLEU の測定には SacreBLEU [8] を用い，2.3 節と同様に，モデルによって生成した理由または結論と，評価データの参照文との間でスコアを計算する．接続表現の品質のばらつきによる影響を抑えるため，8 個の接続表現の候補のうち，開発データでの BLEU が最も高い 5 個をモデルの性能比較に用いる．パープレキシティは，言語モデルの確率推定において，出力すべき文がどれだけ妥当と予測されるかを計測する．

人手評価 人手評価では開発データでの BLEU が最も高い接続表現で推論した理由・結論を評価してもらう．評価者は株式会社サイバーエージェントのアノテータで，出力した理由（または結論）の内容や形式が，モデルに前提として与えた結論（または理由）に対してどれだけ適切か，絶対評価と相対評価をお願いした．評価基準を表 1 に示す．絶対評価では，定められた基準に従って各モデルの出力を評価する．相対評価では，同じ前提文に関して，GPT-2 と RGPT-2 のどちらの出力が優れているかを比較してもらう．相対評価でも絶対評価と同様の基準を採用するが，3 段階の基準では表せないような僅かな差も評価できる．優劣を判断できない場合は「わからない」という回答してもらうことにした．

3.2 実験結果

自動評価 表 2 に自動評価の結果を示す（詳細は付録 B 参照のこと）．開発データでの BLEU スコア

4） 日本語として読みやすい順序で記述したが，RGPT-2 は右から左の向きに単語列の予測を行うことに注意されたい．
5） https://rinna.co.jp/
6） https://github.com/rinnakk/japanese-pretrained-models
7） 2021 年 12 月 13 日時点のもの
8） 例えば RGPT-2 に代名詞が含まれる文を入力した場合，逆方向の文生成が持つ特性により，常識推論タスクのみならず，代名詞が指している名詞を復元するタスクが意図せず含まれ，文生成が不当に難しくなる可能性がある．

表 1　人手評価の基準

	内容の評価	形式の評価
低	どのような状況を想定しても、 理由（結論）とはならない	全く理由（結論）を表す形式の文に なっていないか、文が崩れている
中	結論と理由の文からは直接読み取れない状況を 想定することで、適切な理由（結論）となる	不要な言葉がついていたり、 必要な言葉が脱落していたりすることで、 理由（結論）を表す文の形式として完全ではない
高	何らかの想定をせずとも、 理由（結論）が妥当であると言える	理由（結論）を表すのに適切な 形式の文となっている

表 2　自動評価結果

タスク	モデル	BLEU	Perplexity
理由の推論	GPT-2	0.80	**23.01**
	RGPT-2	**0.90**	23.70
結論の推論	GPT-2	**2.58**	24.86
	RGPT-2	0.90	**17.00**

表 3　理由推論の人手評価結果 [%]

評価軸	モデル	低	中	高	相対
内容	GPT-2	40.7	40.1	19.2	32.4
	RGPT-2	27.5	32.4	**40.1**	**57.7**
形式	GPT-2	16.5	39.0	44.5	34.6
	RGPT-2	12.6	29.1	**58.2**	**61.0**

表 4　結論推論の人手評価結果 [%]

評価軸	モデル	低	中	高	相対
内容	GPT-2	31.9	39.0	**29.1**	**48.4**
	RGPT-2	44.0	32.4	23.6	39.0
形式	GPT-2	17.6	39.0	**43.4**	**55.5**
	RGPT-2	26.4	50.0	23.6	38.5

表 5　順方向（GPT-2）および逆方向（RGPT-2）の言語モデルの出力例

タスク	入力	モデル	内容	形式	出力
理由の推論	私は運動した。	GPT-2	×	×	昨日の夜から腰が痛くて動けなかったからだ。
		RGPT-2	○	○	主治医の先生から「運動しなさい」と言われていたからだ。
結論の推論	私は 1 日中勉強した。	GPT-2	○	○	自分でも驚くほどの点数を取ることができた。
		RGPT-2	×	×	私は今まで、資格試験の勉強をしても意味がないと思っていた。

がトップ 5 に入る接続表現のみで，評価データにおける評価値の平均をとったものを比較する．BLEU（値域は 0〜100）による評価では，理由の推論では RGPT-2 が，結論の推論では GPT-2 が優れており，仮説通りの結果となっている．ただ，いずれも BLEU スコアの絶対値は低い．これは理由または結論に対して様々な説明が可能であることに起因すると考えられる．

一方，パープレキシティではその優劣が逆転している．パープレキシティの評価では理由と結論の両方がモデルに与えられるため，このタスクにおけるモデルの文生成能力を評価するという目的に合致しないことが原因と考えられる．

人手評価　自動評価における BLEU スコア（表 8,9,10,11）により，人手評価で使用する接続表現を決定した．図 1 のように，GPT-2 による理由の推論では「というのも、」，結論の推論では「。結果として、」，RGPT-2 による理由の推論では「ので、」，結論の推論では「なぜなら、」を使用した．

表 3 と 4 に，人手評価の結果を示す．絶対評価で「高」と評価された割合，および相対評価の結果によると，本研究の仮説通り，理由の推論は RGPT-2 が，結論の推論は GPT-2 が優れている結果となり，本タスクを解くことに関する性能に差が表れることを確認できた．すなわち，通常の左から右の言語モデルではなく，文生成向きが右から左となっている逆方向の言語モデルの方が解きやすいタスクが存在することが示唆される．

出力例　モデルによる実際の出力例と相対評価の結果を表 5 に示す．「私は運動した。」という結論に対して GPT-2 による「腰が痛くて動けない」という理由は適切でなく，RGPT-2 の方が適切に理由付けできている．また，「私は 1 日中勉強した。」という理由に対する結論としては，「良い点数を取れた」という GPT-2 の出力内容が自然である．RGPT-2 は，GPT-2 に比べて長い文を出力する傾向にあった．

4　おわりに

日本語における説明の順序の偏りが大規模言語モデルに与える影響を検証するため，単語列の生成向きが順方向あるいは逆方向の大規模言語モデルを構築し，生成的な常識推論タスクに適用した．実験の結果，結論に対して理由を生成するタスクでは逆方向言語モデルが優れており，理由に対する結論を生成するタスクでは順方向言語モデルが適していることが分かった．これらのタスクにおける両モデルの性能差は，日本語における，事柄を記述する順序の偏りが言語モデルに引き継がれているためだと推測される．

本稿では，これまでまだ利用が十分に検討されてこなかった，単語列の生成の向きに関して逆方向の言語モデルの価値を示した．今後は，日本語以外の言語での実験や，常識推論以外のタスクでの検証を進めていきたい．

謝辞

本研究にあたってご助言いただいた，株式会社サイバーエージェントの皆様に感謝いたします．

参考文献

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[2] 影山太郎. 動詞意味論 : 言語と認知の接点. 日英語対照研究シリーズ, No. 5. くろしお出版, 1996.

[3] Peter West, Ximing Lu, Ari Holtzman, Chandra Bhagavatula, Jena D. Hwang, and Yejin Choi. Reflective decoding: Beyond unidirectional generation with off-the-shelf language models. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 1435–1450, Online, August 2021. Association for Computational Linguistics.

[4] 山田康輔, 人見雄太, 田森秀明, 岡崎直観, 乾健太郎. 指定語句を確実に含む見出し生成. 言語処理学会第 27 回年次大会, 2021.

[5] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[6] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In **AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning**, 2011.

[7] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. **CoRR**, Vol. abs/1609.08144, , 2016.

[8] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics.

A　接続表現の出現回数

本研究で用いている各接続表現が、言語モデル学習コーパス内に出現した回数を表 6 と 7 に示す．「(理由)ので、(結論)」のように，順方向に理由から結論を導く接続表現の方が，結論から理由を導く接続表現よりも多いことが分かる．ただし，表に示した接続表現の出現回数は，理由と結論を接続する用法以外の表現も含む場合があるため，注意が必要である．

表 6　順方向に結論から理由を導く接続表現

接続表現	出現回数
というのは、	1,271,177
なぜなら、	228,365
だって、	223,764
というのも、	180,033
その理由は、	56,412
それもそのはず、	17,245
なぜかと言うと、	7,048
この理由は、	2,784
合計	1,986,828

表 7　順方向に理由から結論を導く接続表現

接続表現	出現回数
ので、	33,984,642
から、	11,162,021
ために、	2,458,165
。そのため、	471,595
。これにより、	66,372
。このため、	62,282
。結果として、	10,032
。これを受けて、	2,467
合計	48,217,576

B　自動評価の結果の詳細

表 8,9,10,11 に自動評価の結果の詳細を示す．使用する接続表現に依存して，言語モデルによる理由・結論の推論の性能が大きく変化することが分かる．

表 8　GPT-2 による理由の推論の自動評価

	開発		評価	
接続表現	BLEU	Perplexity	BLEU	Perplexity
というのも、	**1.2**	20.87	0.7	21.48
なぜかと言うと、	1.1	19.73	0.7	20.20
それもそのはず、	1.0	21.52	1.2	21.75
なぜなら、	0.9	25.36	0.8	26.30
その理由は、	0.9	24.28	0.6	25.34
この理由は、	0.7	28.74	0.8	30.24
というのは、	0.6	21.25	0.6	21.78
だって、	0.4	28.90	0.6	29.44
Top-5 平均	1.02	22.35	0.80	23.01

表 9　RGPT-2 による理由の推論の自動評価

	開発		評価	
接続表現	BLEU	Perplexity	BLEU	Perplexity
ので、	**1.6**	20.37	0.9	21.92
から、	1.1	28.55	0.7	31.24
。そのため、	0.9	21.93	0.9	23.33
。結果として、	0.9	18.80	0.7	19.72
。これにより、	0.9	21.24	1.3	22.30
。このため、	0.6	23.93	0.4	25.55
。これを受けて、	0.6	20.52	0.3	21.79
ために、	0.6	23.51	0.8	25.46
Top-5 平均	1.08	22.18	0.90	23.70

表 10　GPT-2 による結論の推論の自動評価

	開発		評価	
接続表現	BLEU	Perplexity	BLEU	Perplexity
。結果として、	**3.5**	20.84	2.6	21.80
。このため、	3.3	26.40	2.1	28.08
。これにより、	3.3	23.29	2.9	24.46
。これを受けて、	3.2	23.24	2.7	24.59
。そのため、	2.9	23.86	2.6	25.38
から、	2.2	27.72	1.0	29.99
ために、	2.1	24.71	0.9	26.33
ので、	1.1	20.45	0.7	21.76
Top-5 平均	3.24	23.53	2.58	24.86

表 11　RGPT-2 による結論の推論の自動評価

	開発		評価	
接続表現	BLEU	Perplexity	BLEU	Perplexity
なぜなら、	**1.4**	16.65	0.9	17.90
だって、	1.4	20.85	0.8	22.45
というのは、	1.3	14.69	0.7	15.68
なぜかと言うと、	1.2	13.45	1.3	14.30
というのも、	0.9	13.86	0.8	14.69
それもそのはず、	0.8	15.77	0.8	16.65
この理由は、	0.7	20.82	0.6	22.64
その理由は、	0.6	17.30	0.7	18.72
Top-5 平均	1.24	15.90	0.90	17.00