早押しクイズにおける超次単語予測の認知モデリング

山下陽一郎　原田宥都　大関洋平
東京大学
{yamashita-yoichiro416,harada-yuto,oseki}@g.ecc.u-tokyo.ac.jp

概要

人間は続きの単語を予測しながら文を理解していると考えられているが、特殊な状況下では、より長いまとまりを持った単語列を予測することができる。本研究では、人間が文処理の際に見せるそのような「超次単語予測」を言語モデルがモデリングできるかどうかを検証する。早押しクイズの問題文を読んでその続きを予測する際の視線を計測し、言語モデルによって、その読み時間をモデリングした。その結果、言語モデルは人間の「超次単語予測」をある程度モデリングでき、さらに特定のデータで言語モデルをファインチューニングすると、よりよくモデリングできることが示された。

1　はじめに

単語の出現確率と人間にとってのその語の認知負荷には関係があるとされ、言語モデルの人間らしさを評価する研究においては、人間が文を読んで理解する際の読み時間や脳活動といった行動データと、言語モデルのサプライザルの間の相関を計算している。[1], [2], [3], [4]

人間の文の読み時間のモデリング研究においては、移動窓式の自己ペース読文実験を行ったり、読み時間のデータのコーパスを用いたりすることが多い。読み時間のデータのコーパスには、Dundee Corpus[5] や Natural Stories Corpus[6] などがあり、日本語では BCCWJ-EyeTrack[7] といったコーパスが存在する。これらのコーパスでは、新聞や小説の文を題材として、人間が文を読んで理解する際の読み時間を計測している。

このような日常的な文理解において、人間が文を読む際には、次の単語を予測しながら文を読んでいることが心理言語学においても示されている。特に、文脈上の意味的な制約は、次単語の予測の助けとなることが示されている。[8]

しかし、人間は文を理解する際、時に次の単語の予測にとどまらないような文全体の予測 (以下、「超次単語予測」と呼ぶ) を働かせることができる。[9] 「超次単語予測」のためには、次に続く文の構造も予測する必要があり、このような「超次単語予測」について焦点を当てたモデリングの研究は少ない。本研究では、文の構造上の制約を手掛かりとして「超次単語予測」を行っていると考えられる早押しクイズに注目して、その問題文を読む際の読み時間のモデリングを行う。

2　サプライザル理論

サプライザル理論 [1], [2] によれば、人間は文処理の際、それ以前の文脈から次に続く語や文節を予測しているとされ、予測しやすい語や文節は認知的負荷が低く、予測しにくい語や文節の認知的負荷が高くなるとされる。さらに、その語や文節の予測しやすさは、(1) のように定式化できるとされる。

$$-\log P(word|context) \tag{1}$$

言語モデルの「人間らしさ」を評価するために、言語モデルが算出したサプライザルと、眼球運動や脳波などの人間から得られるデータを比較する研究が行われてきている。眼球運動データを用いる本研究では、「人間らしい」言語モデルの算出するサプライザルは、各単語の人間の読み時間とよりよく相関すると期待できる。

3　実験

3.1　視線計測実験

実験参加者　本研究の視線計測実験には、32 人の日本語母語話者が参加した。そのうち、7 人が部活やサークルなどで定期的に早押しクイズをしていた経験があり (**expert**)、25 人がそのような経験のない人 (**novice**) であった。

アイテム　本研究では、JAQKET[10] という早押

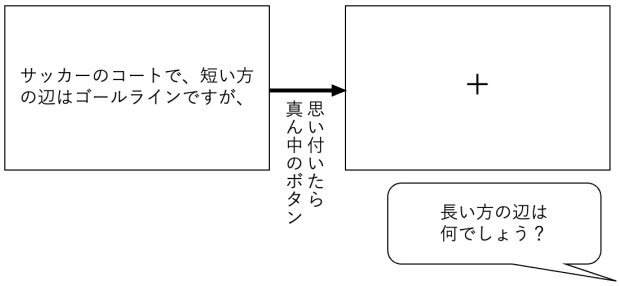

図 1　文産出課題: +predic 条件
パラレルの問題文の後半部を予測する課題

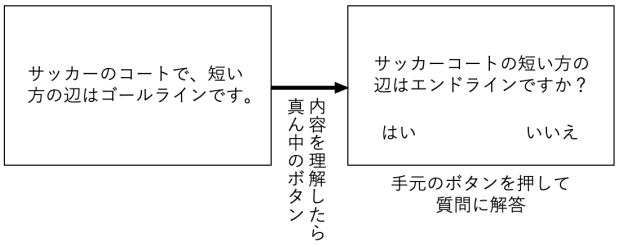

図 2　文理解課題: -predic 条件
パラレルの問題文の前半部を読んで理解する課題

しクイズの問題文のコーパスに含まれるパラレルの問題文を刺激文として用いた。パラレルの問題文は「ですが問題」とも呼ばれ、例えば

> おおいぬ座のアルファ星はシリウスですが、こいぬ座のアルファ星は何？（答. プロキオン）

のように、「A は X ですが、A' は何？」という構造を持つ。パラレルの問題では、問題の本題は後半部に置かれるが、場合によっては問題文の前半のみを聞けば、その後半部を予測できる。本研究では、後半部の予測が容易な問題 (**easy**) と難しい問題 (**difficult**) に区別し、20 問ずつアイテムに含めた。

本研究で **easy** に分類されるパラレルの問題文には以下のようなものがあり、このような問題では、問題文の前半部を読んで問題文の後半部を予測することが可能である。

> フランスの国会で、上院に相当するのは元老院ですが、（下院に相当するのは何でしょう？）

一方、**difficult** に分類されるような次の問題では、問題文の続きを一通りに予測することが困難である。

> スマホで使われる用語で、「アプリ」といえば「アプリケーション」の略ですが、（「アプデ」といえば何の略でしょう？）

課題　本実験では文産出課題 (**+predic**) と文理解課題 (**-predic**) との 2 種類の課題を行なった。

文産出課題においては、パラレルの問題文の前半部を提示し、その続きの文を考える際の眼球運動を計測し、前半部の各単語の読み時間を計測した。

文理解課題においては、パラレルの問題文の前半部を「〜です。」という形に直して提示し、その間の各単語の読み時間を計測した。続く画面では文理解課題を実施した。

3.2　言語モデル

Huggingface 上で rinna 社の公開している [11]GPT-2[12] を用いて、言語モデルのサプライザルを算出した。事前学習済みのモデルと、ファインチューニングを施したモデルの両方を用いて実験を行った。

GPT-2　事前学習のみを行った GPT-2 に、視線計測実験で提示した文を入力した際の、各単語の予測確率 P(単語 | 文脈) に基づいて、各単語のサプライザル $-\log P$(単語 | 文脈) を計算した。GPT-2 は rinna/japanese-gpt2-medium のモデルを使用した[1)]。

ファインチューニング　JAQKET[10]、Quiz-Works[2)]、クイズの杜[3)]で二次使用可能として公開されている早押しクイズの問題文のうち、「ですが、」「ますが、」を含むパラレルの問題 4100 問を抽出し、GPT-2 のファインチューニングを行った。その際、ファインチューニングに用いるデータ数を 10, 100, 200, 300, 500, 700, 1000, 1500, 2000, 4100 の 10 段階に分け、それぞれシード値を変えて 5 回ずつ実験を行った。データ数が 2000 以下の各条件ではファインチューニングに用いるデータをランダムに抜き出した。

3.3　評価指標

各条件の言語モデルについて、算出されるサプライザルの眼球運動のモデリング精度を評価するための指標として、心理学的予測精度 (Psychometric Predictive Power: PPP) を用いる。PPP は、眼球運動をモデル化するベースラインの回帰モデルに、各条件の言語モデルのサプライザルを加えた際の対数尤度 (ΔlogLik) を用いる。先行研究 [4] を参考に、眼球運動のベースライン回帰モデルには、以下の線形混

1)　https://huggingface.co/rinna/japanese-gpt2-medium
2)　https://quiz-works.com/
3)　https://quiz-schedule.info/quiz_no_mori/data/data.htm

表 1　本研究のモデリングに用いた特徴量

変数名	型	概要
length	int	トークンの文字数
is_first	factor	行内最左要素
is_last	factor	行内最右要素
lineN	int	画面内提示行数
segmentN	int	行内提示トークン数
log_freq	num	トークン頻度の対数
prev_length	int	直前トークンの文字数
log_freq_prev	num	直前トークンの頻度の対数
subject_id	factor	被験者 ID
item_id	factor	アイテム ID

合モデルを用いる[4]：

log(TRT) ~ length + is_first + is_last + lineN
+ segmentN + log_freq + prev_length
+ log_freq_prev + (1|subject_id)+ (1|item_id)

単語の出現頻度 (log_freq, log_freq_prev) は、早押しクイズの問題文 87,467 問中における当該トークンの出現頻度を計算した。数値型の特徴量は全て中心化を行った。その他、各特徴量の詳細は表 1 に記す。

4　結果と考察

4.1　事前学習のみの GPT-2

ファインチューニングを施していない、事前学習のみの GPT-2 の PPP を表 2 に示す。

文理解課題 (-predic) と文産出課題 (+predic) で比較すると、文産出課題の際の読み時間の方がより良くモデリングできている。文産出課題では、実験参加者はパラレルの問題文の前半部を読み、問題文の後半部の「超次単語予測」を行っている。このことから、事前学習のみの言語モデルでも、早押しクイズの際に見られるような「超次単語予測」をある程度行っていると考えられる。また、文産出課題 (+predic) における、クイズ経験者 (expert) とクイズ未経験者 (novice) の眼球運動を比較すると、expert の視線の方がより良くモデリングできることがわかった。さらに、問題文後半部の予測のしやすさで比べると、difficult の対比よりも、easy の対比の問題の方が、より良くモデリングできているとわかった。

expert の方が novice よりもパラレルの問題文の形式により慣れており続きの問題文の予測もしやすいこと、そして、easy な対比の問題の方が difficult な対比の問題よりも続きが予測しやすいことを考えると、人間が「超次単語予測」を働かせているような条件であるほど、言語モデルの PPP も高くなることがわかる。言語モデルも「超次単語予測」をある程度モデリングできるということが示された。

4) GPT-2 で使用されているトークナイザによるトークン単位で、各トークンの総読み時間 (Total Reading Time, TRT) を計測した。

表 2　事前学習のみの GPT-2 でのモデリングの結果

条件	データ数	Δ logLik ($/10^3$)	χ^2	P
-predic	7869	0.01493	0.235	0.6278
+predic	8361	**1.489**	24.893	0.0000 ***
+predic novice	6351	0.5396	6.8545	0.008842 **
+predic expert	2010	**1.756**	6.7061	0.009608 **
+predic easy	4579	**1.672**	15.316	0.0001 ***
+predic difficult	3782	0.5577	4.2187	0.03998 *

4.2　ファインチューニングを施した GPT-2

ファインチューニングした後の言語モデルでのモデリングの結果を示したのが図 3 である。パラレルの問題文を用いてファインチューニングした各モデルについて、PPP を縦軸に、そのモデルの Perplexity を横軸にとってプロットした。ファインチューニングに用いたデータ数は 4100, 2000, 1500, 1000, 700, 500, 300, 200, 100, 10, 0 の 10 条件であり、用いたデータ数が多いものほど、大きくプロットされている。なお、一番小さな点でプロットされているのは、ファインチューニング前の事前学習のみのモデルの値である。

文産出課題 (+predic) の expert と novice それぞれが、easy の問題と difficult の問題を読む際の読み時間についてモデリングしているが、どの条件についても、ファインチューニングに用いるデータ数を増やすほど、Perplexity が下がる傾向にあることがわかる。

また、どのデータ数でのファインチューニングにおいても、PPP が最も高いのは、expert が easy な対比の問題文を読む際の眼球運動である。novice の眼球運動は、easy の問題でも difficult の問題でも、ほとんど変わらない PPP となっていた。expert が easy の問題を読むときの条件において、最も予測が強く

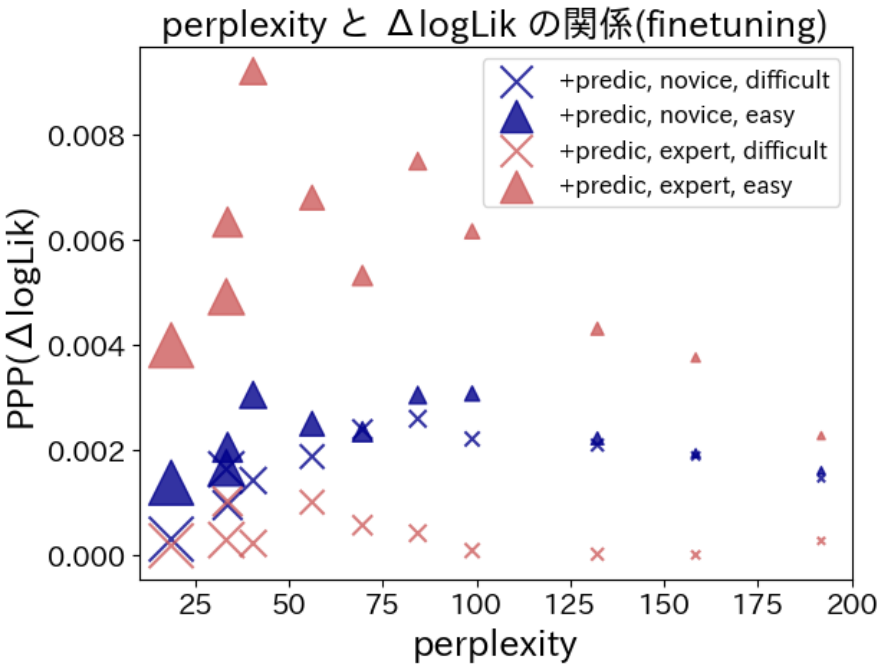

図 3　ファインチューニングした GPT-2 でのモデリング
データ数は大きい方から順に 4100, 2000, 1500, 1000, 700, 500, 300, 200, 100, 10, 0

働くということを考えると、言語モデルも「超次単語予測」を行うことができると示唆される。expert が difficult の問題を読む際は、問題文の続きが一通りに定まらないために「超次単語予測」を働かせることが難しいが、そのような問題でも novice は決めうち的に問題文の続きを予測しているためある程度の予測が働き、その結果、novice, difficult の条件の方が、expert, difficult の条件よりも PPP が大きくなったのではないかと考える。

どの条件の読み時間のモデリングにおいても、ファインチューニングに用いるデータ数を増やすと、PPP は増加していくが、データ数が 1000 を超えると、一転して PPP が減少していくような傾向が見られる。多くのデータでファインチューニングをすると次単語予測が簡単になり、サプライザルが全体的に下がり過ぎてしまうため、読み時間との相関が下がることになると考えられる。日本語の読み時間のモデリングにおいて、このように Perplexity が下がり過ぎてしまうと、PPP も減少に転じるという傾向は先行研究 [4] で示されている結果とも一致する。

5　おわりに

本研究では、人間の文処理の際の「超次単語予測」に焦点を当てるため、早押しクイズの問題文を読む際の眼球運動を計測し、その読み時間のモデリングを行った。結果としては、事前学習のみの GPT-2 でも、ある程度の精度でモデリングすることができ、言語モデルも「超次単語予測」のモデリングが可能だと示された。

また、早押しクイズの問題文を用いて言語モデルにファインチューニングを施すと、さらに PPP が向上していくことがわかった。特に、+predic, expert, easy 条件という「超次単語予測」が一番顕著に見られるだろう条件において最も PPP が高く、ファインチューニングによって PPP の値が目立って増加した。このことからも、言語モデルは人間の「超次単語予測」をモデリングすることは可能であり、ファインチューニングによってその精度が向上することがわかる。しかし、ファインチューニングに用いるデータ数が 1000 を超えると、全体的にサプライザルが下がり過ぎてしまうことにより PPP が減少に転じるという、先行研究と合致する傾向が見られた。

パラレルの問題の予測には、どこを強調して読まれるかといった音韻的な要素も大きな手がかりとなっていることが指摘されており、音韻的な情報も取り入れた予測処理については今後の研究での課題として残されている。

6 謝辞

本研究は、JST さきがけ JPMJPR21C2 および JSPS 科研費 JP20H01254 の支援を受けたものです。

参考文献

[1] John Hale. A probabilistic Earley parser as a psycholinguistic model. In **Second Meeting of the North American Chapter of the Association for Computational Linguistics**, 2001.

[2] Roger Levy. Expectation-based syntactic comprehension. **Cognition**, Vol. 106, No. 3, pp. 1126–1177, 2008.

[3] Nathaniel J Smith and Roger Levy. The effect of word predictability on reading time is logarithmic. **Cognition**, Vol. 128, No. 3, pp. 302–319, 2013.

[4] Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. Lower perplexity is not always human-like. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **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. 5203–5217, Online, August 2021. Association for Computational Linguistics.

[5] Alan Kennedy and Joël Pynte. Parafoveal-on-foveal effects in normal reading. **Vision research**, Vol. 45, No. 2, pp. 153–168, 2005.

[6] Richard Futrell, Edward Gibson, Harry J Tily, Idan Blank, Anastasia Vishnevetsky, Steven T Piantadosi, and Evelina Fedorenko. The natural stories corpus: a reading-time corpus of english texts containing rare syntactic constructions. **Language Resources and Evaluation**, Vol. 55, pp. 63–77, 2021.

[7] Masayuki Asahara, Hajime Ono, and Edson T. Miyamoto. Reading-time annotations for "Balanced Corpus of Contemporary Written Japanese". In **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers**, pp. 684–694, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[8] Susan F Ehrlich and Keith Rayner. Contextual effects on word perception and eye movements during reading. **Journal of verbal learning and verbal behavior**, Vol. 20, No. 6, pp. 641–655, 1981.

[9] 伊沢拓司. クイズ思考の解体. 朝日新聞出版, 2021.

[10] 鈴木正敏, 鈴木潤, 松田耕史, 西田京介, 井之上直也. Jaqket: クイズを題材にした日本語 qa データセットの構築. 言語処理学会第 26 回年次大会, pp. 237–240, 2020.

[11] 趙天雨, 沢田慶. 日本語自然言語処理における事前学習モデルの公開. 人工知能学会研究会資料 言語・音声理解と対話処理研究会 93 回 (2021/11), pp. 169–170. 一般社団法人 人工知能学会, 2021.

[12] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1, No. 8, p. 9, 2019.