情報量に基づく日本語項省略の分析

石月由紀子 [1]　栗林 樹生 [1,2]　松林 優一郎 [1,4]　大関 洋平 [3,4]
[1] 東北大学 [2]Langsmith 株式会社 [3] 東京大学 [4] 理化学研究所
yukiko.ishizuki.p7@dc.tohoku.ac.jp,
{kuribayashi,y.m}@tohoku.ac.jp, oseki@g.ecc.u-tokyo.ac.jp

概要

日本語は英語などの言語と比べて項省略が生じやすい言語である．本研究では，人がどのような基準で日本語の項省略を行うか，書き手の計算モデルについて探求する．具体的には，「書き手は読み手にとっての処理負荷を想定しながら文を書いており，後続要素の処理負荷をなるべく大きくしないように省略を行っている」という仮説を立てる．項に後続する単語系列の処理負荷の近似として言語モデルが計算するサプライザルを用いて分析を行い，結果として，後続する単語の処理負荷 (サプライザル) が大きくならないよう省略を行っているという，我々の仮説を支持する結果が得られた．

1　はじめに

言語を用いた情報伝達における主要な目的の一つは，情報を相手に正確に伝えることである．しかし，情報を正確に伝えるという目標に反して日本語では文要素の省略現象が頻繁に観察され，なぜ・どのようなときに省略が生じるのかは言語学的関心を集めてきた．どのような要素が省略可能・不可能であるかといった説明は長らく生成統語論で議論されてきたが [1, 2, 3, 4]，言語運用における実際の書き手の判断 (省略する/しない) の傾向については分析が限られている．

本研究ではこのような書き手の省略判断の選好について，情報理論に基づくアプローチで説明を試みる．情報理論的，確率的な説明は，省略や指示表現選択といった言語産出現象の説明としばしば相性がよく，これまでにも発話中の情報量の分布がなめらかになるような情報伝達を仮定する情報密度一様性 (UID) 仮説 [5, 6] や，人間の読み負荷とサプライザルの関係 [7] に基づいて，関係詞省略 [6] や縮約 [8], 格助詞の省略 [9, 10] などが分析されてきた．

本研究では，「書き手が読み手にとっての処理負荷を想定しながら文を書いており，後続要素の処理負荷をなるべく大きくしないように省略を行っている」という仮説を立てる．読み手にとっての処理負荷の増大と，要素を省略しないことによる文の長さ (冗長性) にはトレードオフがあり，両方の要請を満たすには，処理負荷を増大させない範囲で省略を行うという戦略が妥当であると考えられる．例えば，「家のコーヒー豆が切れたので、デパートでコーヒー豆を買った」という文については，述語動詞の項 (コーヒー豆を) が省略されていても無理なく後続の情報 (買った) の解釈が可能であることから省略ができると考える．

実験では，動詞の項の省略に焦点を当て，項の省略とその省略によって引き起こされる後続文脈の処理負荷の変化の関係を調査した．後続文脈の処理負荷については，サプライザル理論に則り，累積サプライザルの変化 (項が省略されたときに，後続要素の驚きがどれほど増えるか) を観察する．述語項構造・共参照アノテーション付きデータを用い，テキスト上で省略されている項といない項について，その項の表出・省略によって引き起こされる後続単語の処理負荷の変化量を調査し，「後続する要素の処理負荷が増えないように省略を行っている」という仮説を支持する結果が得られた．

2　サプライザルによる処理負荷推定

人の読み処理や言語の産出過程については，計算心理言語学の視点から分析が行われている．特に近年では，逐次的な単語の処理困難度と情報量の間の相関を説明するサプライザル理論 [7] が，読み時間のモデリング [11, 12] などに応用されている．サプライザルは, 先行文脈 $c = w_1, ..., w_{i-1}$ に続く単語 w_i の出現確率について負の対数をとったものであり，

$$\mathtt{surprisal}(w_i \mid c) = -\log_2 P(w_i \mid c) \tag{1}$$

によって求められる．日本語においても，読み時間の長さとサプライザルの大きさに相関があることが示されている [13]．さらには，共参照 [14] やフラグメント [15] といった言語現象についても，サプライザルを用い情報量の観点で分析が与えられている．本研究では，日本語の項省略について後続の語の読み処理負荷との関係を観察するために，サプライザルを用いて分析を行う．

3 方法

本研究での仮説は「書き手は読み手にとっての処理負荷を想定しながら文を書いており，日本語において自然に省略される項は，項の出現の有無によって引き起こされる後続単語の処理負荷の変化量が相対的に小さい」というものである．この仮説を検証するため，(1) まず予備実験的に，文章中に登場する述語の項を「自然に省略される項」と「自然には省略されない項」の 2 つのグループに分け，それぞれのグループで項を出現させた場合と項を出現させない場合を比べ，後続の語のサプライザルにどの程度の変化が現れるかを確認する．(2) その後，項省略に関連する可能性のあるその他の因子を交えて回帰分析を行い，実際に後続の語のサプライザルの変化量が書き手が項省略を起こすか否かの予測に寄与しているかどうかを検証する．

人手による実験用データの作成　上記の検証を行うにあたっては，書き手が項を明示している事例と省略を行っている事例を区別したデータを作成する必要がある．本研究では係り受け．並列構造と述語項構造・共参照がアノテーションされた現代日本語書き言葉均衡コーパス (BCCWJ-DeparaPAS)[16] の書籍（PB）ドメインから述語項関係を持つ事例を抽出することでこのデータを作成した．

作成したデータの概略図を図 1 に示す．まず，文章中の述語の項について，書き手が項を明示している事例を **(A) 項表出群** (付録 表 3) とし，省略を行っている事例を **(B) 項省略群** (付録 表 4) とした．ガ，ヲ，ニ格の 3 つの表層格を対象とし，(A) 群，(B) 群共にそれぞれの格に対し，20 事例ずつ，計 120 の事例をサンプルした．

さらに，この事例に対して項の出現の有無による後続の語のサプライザルの変化を計算するために，**(A) 項表出群**については，明示されている項を削除した文を新たに作成した．一方で，**(B) 項省略群**については，省略された項を文内に補った文を作成した．この場合の項の挿入位置は著者ら 3 名で合議の上，アノテーションされている述語項構造の意味が保たれてかつ，もっとも日本語として自然な文となる位置に定めた．作成の詳細な手順については付録 A に記した．

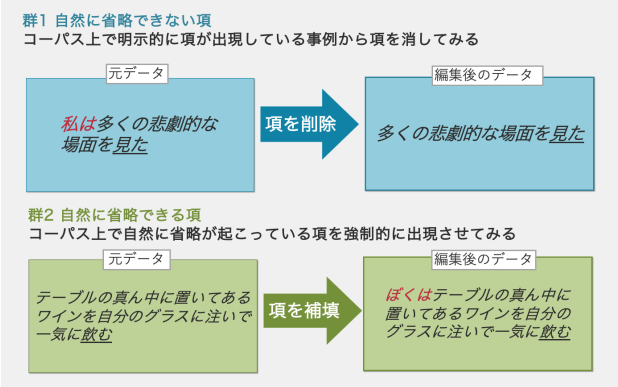

図 1　(A) 項表出群と (B) 項省略群における後続文脈のサプライザル比較の概略図．(A) 群，(B) 群共に項が表出している場合と表出していない場合の文を作成し，項以降のサプライザル差を比較する

4 実験

4.1 実験設定

言語モデル　サプライザルの計算には，新聞と日本語 wikipedia で学習した Transformer ベースの left-to-right 言語モデルの単語予測確率を使用した [17]．3 節で作成した事例について，ターゲットとなる述語より前方 300 サブワードを前方文脈とし，前方文脈と対象述語を含む文をつなげたものを入力系列とした．直前の語までの入力に基づいて計算される次の単語の予測確率を用いて項の出現位置以降の各単語のサプライザルを計算した[1)]．言語モデルの詳細は付録 D に示す．

4.2 予備実験

項の有無によるサプライザルの変化　図 2 (1) は，文中のある述語の項を仮に出現させなかった場合の後続する単語列の処理負荷から，出現させた場合の処理負荷を引いた値について，その傾向を示したものである．この値が大きいほど，項を出現させな

1)　言語モデルの入力には，入力文を mecab と unidic で国語研短単位に分割し，それをさらに BPE (Bite Pair Encoding) によってサブワード化したものを用いた．サプライザルの計算時は国語研短単位ごとにサブワードのサプライザルの和を取った．

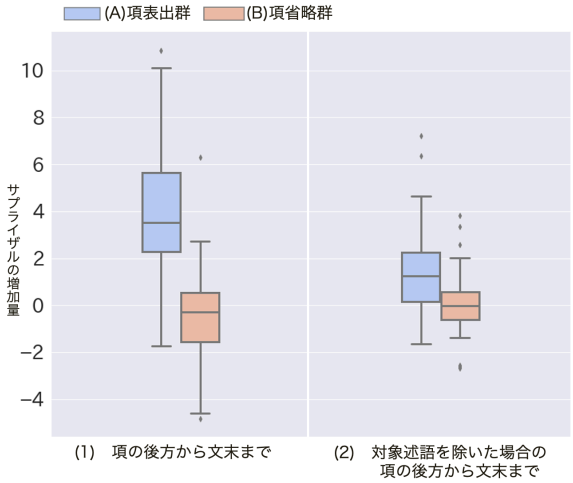

図 2 (1) は，項を表出させた時とさせないときの，項以降の単語のサプライザルの総和の変化を示す．(2) は，項以降の単語のサプライザルの総和の変化のうち，対象の述語動詞のサプライザルの変化を除いた値を示す．(1) における変化の大きさに比べ，(2) の変化の大きさが小さいことから，項を表出させるときの後続文脈の情報量の変化は主に対応する動詞で起きていることがわかる．

かった場合に，後続する情報の処理負荷が増大することを意味する．後続する単語列の処理負荷としては，具体的に項から文末までの単語のサプライザルの総和を用いた．

テキスト上で省略されている項は省略しても後続する処理負荷が増大しにくい 図中で，(A) 項表出群は書き手が実際に項を表出させた群であり，(B) 項省略群は書き手が実際に項を省略させた群である．図 2 (1) について (A) 群と (B) 群の結果を比較すると，(A) 群の中央値が 2.24，(B) 群の中央値が −0.12 と，項表出群のサプライザル差が項省略群より相対的に大きい値となった．また，この傾向はすべての表層格で同様に観測できた (付録 C を参照)．つまり，書き手が省略を選択した群の方が，省略を選択しなかった群に比べて，省略を行うことによって発生する後方単語のサプライザルの増加量が相対的に小さくなっている．この結果は「後方の単語の読み処理負荷 ($\propto$ サプライザル) を増大させないように書き手が省略する項を選択している」という仮説を支持している．

処理負荷の増大は主に動詞で生じている 具体的に後方のどの部分で主たるサプライザルの変化が起こっているかを観察したところ，事例の多くでターゲットの動詞部分で相対的に大きな変化が見られた．図 2 (2) は，後続文脈の処理負荷 (サプライザル変化の総和) のうち，対応する動詞のサプライザル変化を除いた量を示す．項の後方から文末まで単語のサプライザル変化の総和 (図 2 (1)) のうち，対象の述語動詞以外の部分のサプライザルの変化 (図 2 (2)) は小さい．このことから，項の有無は後述の単語の中でも特に述語の処理負荷により大きな影響を与えていることが示唆され，述語のサプライザルが項省略の可否に関わることが予想される．このことをより精緻に分析するため，次節では回帰分析を行う．

4.3 回帰分析

前節 4.2 で示された述語のサプライザル差の影響について，項省略に関連する可能性のあるその他の因子を加えて回帰分析を行い，実際に後続の語のサプライザルの変化量が，書き手が項省略を起こすか否かの説明に寄与していることを確かめる．具体的には，述語動詞のサプライザル差以外に，以下の 6 つの因子を考慮し，コーパス上において書き手が項を省略させていたかどうかを目的変数とした回帰分析を行う．[2)]

省略可能性に影響すると想定されるものとして，項の情報状態に関連する素性を 2 つ選択した．まず，先行研究 [13] より，名詞句の情報状態に伴う読み負荷の傾向がサプライザルでも再現されることが確認されていることから，項が新情報か旧情報かを近似的に表現する素性として (1) 項名詞句のサプライザルを用いた．加えて，(2) 項名詞句と共参照関係にある名詞句が項の挿入位置よりも前に存在するかどうかを素性とした．さらに，述語項構造解析で使用される基本的特徴量として，(3) 項名詞句の長さ，(4) 項を含む文の長さ，(5) 項が何文目に出現するか，(6) 項と述語の単語距離を選定した (表 1)．

省略可能性とサプライザル 構築したモデルの回帰式を式 2 に示す．

```
dep_zero ~ diff_verb_spr + arg_spr
         + l_arg + arg_eq + l_sent      (2)
         + appearance + arg2verb
```

推定結果を表 2 に示す．選択された説明変数のうち，有意水準 5%において帰無仮説を棄却し，有意性が認められた素性は，(a) 述語動詞のサプライザル差，(b) 項のサプライザル，(c) 項と述語の単語距離の 3 つであった．(a) より，他の素性を考慮してもなお，述語動詞のサプライザルの変化が小さいほど

2) これらの素性はステップワイズ法によって選定し，素性間に強い相関がないことを確認した．

表 1　回帰分析に用いた素性の一覧

素性	型	摘要
dep_zero	ブーリアン	項がコーパス上で省略されているか
diff_verb_spr	実数	述語動詞のサプライザル差
arg_spr	実数	項名詞句のサプライザル
l_arg	整数	項名詞句の長さ
arg_eq	ブーリアン	項名詞句が共参照名詞句か
l_sent	整数	項を含む文の長さ
appearance	整数	項が何文目に出現するか
former_noun	カテゴリカル	項名詞句が述語以前に出現するか
arg2verb	整数	項と述語の距離

省略されやすいという傾向が示された．(b) については，読み手にとって出現が予測しやすい項ほど省略されやすいという傾向が観察され，旧情報のもつ情報量が小さいことを想定すると，旧情報ほど省略されやすいという言語的直観と一致する．(c) 動詞と離れている項が省略されやすいという傾向については，関連する語同士は近くに配置されるという言語一般的な傾向を踏まえると，述語との関連が弱い単語ほど省略されやすいということを示している．また，主題や主語など，文頭の要素が省略されすいという傾向と一貫する．

5　議論

今回は項と後続要素との関係に着目してサプライザルを比較したが，今回説明変数に組み込まなかった部分においても議論の余地が残されている．例えば，そもそも書き手が通常あまり書かれないような稀な情報を記述したい場合は，その情報を記述すれば記述するほど文のサプライザルは上昇するのであり，その場合，項を省略すればするほど全体のサプライザルは低下することになる．しかしながら，著者が伝えたい情報が真にそれであるならば，情報を省略することはできない．したがって，理想的には，書き手が伝達したい情報が十分に読み手に伝わるという制約条件の元での読み手の処理負荷最小化問題といったような問題の定式化が望まれる．今後は，そのような制約付き問題の分析の第一歩として，分析対象の項が新情報であるか旧情報であるかを明確に区別したデータ上で同様の分析を行うことを検討する．

表 2　回帰分析の結果．$P|z|$ における*の数は，係数が 0 であるという帰無仮説を有意水準 90%(*)，95%(**)，99%(***) で棄却したことを示す．

Parameter	coef	std err	$P > \|z\|$
diff_verb_spr	0.5780	0.192	0.003***
arg_spr	-0.2320	0.100	0.021**
l_arg	0.2884	0.207	0.163
arg_eq	1.1362	1.320	0.389
l_sent	-0.0043	0.020	0.833
appearance	0.0036	0.004	0.351
arg2verb	0.1381	0.033	0.000***

また，「項省略によって大きくなる述語動詞の情報量が，省略によって小さくなる項自体の情報量を上回る場合には省略ができない」という UID の観点からの解釈も興味深く，UID の観点を含めた更なる分析は今後の課題としたい．

6　おわりに

本研究では，読み時間や縮約などの現象に関する既存研究で導入されていたサプライザル理論を日本語項省略の分析に拡張し，情報量の観点から人がどのような基準で項省略を行うかという書き手の計算モデルを探索した．結果として，項の有無によるサプライザルの変化を分析した結果，書き手が省略を選択する項は，省略しないことを選択する項と比べて，省略することで発生する後方の単語の読み処理負荷の変化量が相対的に小さくなるという仮説を裏付ける結果が得られた．また，回帰分析の結果から，省略を行うことで発生する述語動詞のサプライザルの増加量が項省略の因子として存在することが明らかになった．今後は分析対象のデータの規模の拡大や，より多くの素性を考慮した分析を行うことを目標としたい．これに加え，人にとっての「自然な省略」を定量化するため，個々人で省略するかしないかの判断の揺れが生じる可能性のある項の事例について，省略の自然さに関する人間のアグリーメントを取ることを検討している．

謝辞

本研究は，JSPS 科研費 JP19K12112，JP20J22697，および JST さきがけ JPMJPR21C2 の支援を受けたも

のです.

参考文献

[1] S.-Y. Kuroda. Whether we agree or not. **Lingvisticae Investigationes**, Vol. 12, pp. 1–47, 1 1988.

[2] Mamoru Saito. Notes on east asian argument ellipsis. **LANGUAGE RESEARCH**, Vol. 43, pp. 203–227, 2007.

[3] Mamoru Saito. (a) case for labeling: Labeling in languages without φ-feature agreement. **Linguistic Review**, Vol. 33, pp. 129–175, 2 2016.

[4] Serkan Sener and Takahashi Daiko. Argument ellipsis in japanese and turkish. **MIT Working Papers in Linguistics 61 : Proceedings of the 6th Workshop on Altaic Formal Linguistics : Department of Linguistics and Philosophy. MIT**, pp. 325–339, 2010.

[5] T Jaeger and Roger Levy. Speakers optimize information density through syntactic reduction. Vol. 19, pp. 849–856. MIT Press, 2007.

[6] T Florian Jaeger. Redundancy and reduction: Speakers manage syntactic information density. **Cognitive Psychology**, Vol. 61, pp. 23–62, 8 2010.

[7] John Hale. A probabilistic earley parser as a psycholinguistic model. pp. 159–166, 2001.

[8] Austin F. Frank and T. Florain Jaeger. Uc merced proceedings of the annual meeting of the cognitive science society title speaking rationally: Uniform information density as an optimal strategy for language production permalink speaking rationally: Uniform information density as an optimal strategy for language production. **Proceedings of the Annual Meeting of the Cognitive Science Society**, Vol. 30, p. 30, 2008.

[9] Chigusa Kurumada and T Florian Jaeger. Communicative efficiency in language production: Optional case-marking in japanese. **Journal of Memory and Language**, Vol. 83, pp. 152–178, 12 2015.

[10] ELISABETH NORCLIFFE and T. FLORIAN JAEGER. Predicting head-marking variability in yucatec maya relative clause production. **Language and Cognition**, Vol. 8, pp. 167–205, 6 2016.

[11] 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.

[12] Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. Lower perplexity is not always human-like. **ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference**, pp. 5203–5217, 2021.

[13] 栗林樹生, 大関洋平, 伊藤拓海, 吉田遼, 浅原正幸, 乾健太郎. 日本語の読みやすさに対する情報量に基づいた統一的な解釈. 言語処理学会第 27 回年次大会発表論文集, pp. 723–728, 2021.

[14] Laura Aina, Xixian Liao, Gemma Boleda, and Matthijs Westera. Does referent predictability affect the choice of referential form? A computational approach using masked coreference resolution. **CoRR**, Vol. abs/2109.13105, , 2021.

[15] Robin Lemke, Ingo Reich, Lisa Schäfer, and Heiner Drenhaus. Predictable words are more likely to be omitted in fragments–evidence from production data. **Frontiers in Psychology**, Vol. 12, p. 2266, 7 2021.

[16] 浅原正幸, 大村舞. Bccwj-depparapas:『現代日本語書き言葉均衡コーパス』係り受け・並列構造と述語項構造・共参照アノテーションの重ね合わせと可視化. 言語処理学会第 22 回年次大会発表論文集, pp. 489–492, 2016.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In **Proceedings of NIPS**, pp. 5998–6008, 2017.

[18] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In **Proceedings of EMNLP**, pp. 66–71, 2018.

A　検証に用いたデータの作成手法

本研究では係り受け．並列構造と述語項構造・共参照がアノテーションされた現代日本語書き言葉均衡コーパス (BCCWJ-DeparaPAS) [16] の書籍（PB）ドメイン 83 ファイルを用いて検証を行った．コーパスには，名詞句相当の単語に項 id が付与され，述語相当の単語に NTC 形式の属性，及び項との関係が付与されている．ゼロ代名詞については，ga_dep="zero"というようなタグ付けがされている．項の格助詞については，コーパス上における格属性のうち，ガ格，ヲ格，ニ格の 3 パターンを対象とし，述語に関しては本研究では用言述語のみを対象とした．また，用言述語のうち，する，ある，なるのような機能性の強い語については，予め除外した．さらに，各群について項の格パターンに分類した上で，ガ格，ニ格，ヲ格それぞれ 20 サンプルを無作為に抽出した．まず，(A) 項表出群については，コーパス上で表出している項を消去したデータを作成した．続いて (B) 項省略群については，コーパス上で省略されている項を元文に補填したデータを作成した．その際，補填する項の格助詞については，例えば主格「僕が」を「僕は」に変更するような表出形の変化を許容し，その挿入位置は著者ら 3 名で合議の上，元文の述語項構造において自然と感じる位置に補填した．

B　作成したデータの例

表 3　**(A) 項表出群**のデータ作成例
(BCCWJ:00003_A_PB59_00001)

元データの文章	**そこには**八階で降りようとする男が映っていた。
表出している項	そこには
削除された項	そこには
項省略後の文章	八階で降りようとする男が映っていた。

表 4　**(B) 項省略群**のデータ作成例
(BCCWJ:00001_A_PB12_00001)

元データの文章	各駅に着く前に必ず一瞬真っ暗になった昔の地下鉄を思い出す。
省略された項	私+ガ格
補填された項	私は
項補填後の文章	各駅に着く前に必ず一瞬真っ暗になった昔の地下鉄を**私は**思い出す。

C　予備実験の結果

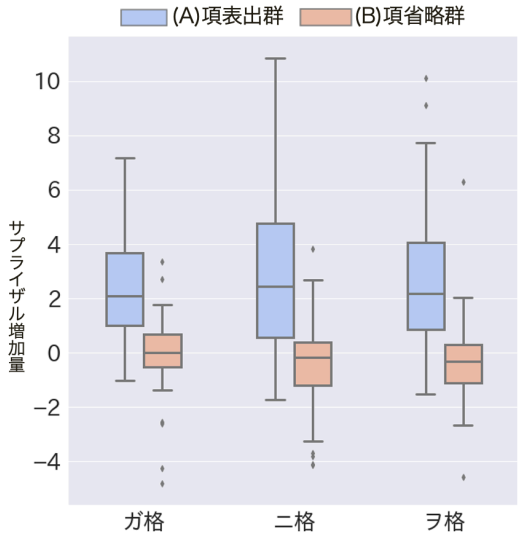

図 3　述語の項を出現させなかった場合の項から文末の単語のサプライザルの総和から，仮に出現させた場合の項以降の単語のサプライザルの総和を引いた値に関する表層格間の比較．

D　言語モデルハイパーパラメータ

Transformer 日本語言語モデルを，500 万文 (日本語 Wikipedia と新聞) で学習し，100,000 回アップデート後のチェックポイントを用いた．日本語テキストは，アノテーションデータとの分割の一貫性を保つため，一度国語研短単位に分割した後，BPE でサブワードに分割した[3)]．

表 5　言語モデルのハイパーパラメータ．

Fairseq model	architecture	transformer_lm_gpt
	adaptive softmax cut off	50,000, 140,000
	share-decoder-input-output-embed	True
	embed_dim	384
	ffn_embed_dim	2,048
	layers	8
	heads	6
	dropout	0.1
	attention_dropout	0.1
Optimizer	algorithm	AdamW
	learning rates	5e-4
	betas	(0.9, 0.98)
	weight decay	0.01
	clip norm	0.0
Learning rate scheduler	type	inverse_sqrt
	warmup updates	4,000
	warmup init lrarning rate	1e-7
Training	batch size	61,440 tokens
	sample-break-mode	none

3)　SentencePiece [18] を用い，文字の網羅率を 0.9995，語彙数を 100,000 とした．