# 翻訳言語モデルを中間タスクとするゼロ照応解析

馬越 雅人　村脇 有吾　黒橋 禎夫
京都大学大学院情報学研究科
{umakoshi, murawaki, kuro}@nlp.ist.i.kyoto-u.ac.jp

## 概要

日本語と非プロドロップ言語の対訳テキストは，前者で省略されるゼロ代名詞が多くの場合後者では明示的に言及されるため，日本語ゼロ照応解析の精度を改善する可能性を秘めている．対訳テキストを活用するために機械翻訳を中間タスクとする手法が提案されているが，事前学習タスクとの差異が大きく，改善の余地がある．本研究では，より事前学習タスクに近い翻訳言語モデルを中間タスクとする手法を提案する．実験により，翻訳言語モデルを用いた場合の方が精度が向上することを示す．

## 1　はじめに

誰が誰に何をしたかは情報の基本的な単位であり，これを理解することは自然言語理解において必要不可欠である．しかし，日本語や中国語のようなプロドロップ言語では，文脈から推測できる代名詞が省略されることがあり，特に困難な問題となる．

(a) ($\phi_i$=ガ) 気が向いたら行くよ。 [著者]

(b) I will go if I feel like.

**図 1**　日本語で省略されたゼロ代名詞が英語では明示されるような対訳文の例．日本語で省略されたガ格の項 (著者) が英語では明示的に現れている (I).

図 1(a) のように，省略されたゼロ代名詞 ($\phi$) の照応先を特定することをゼロ照応解析と呼ぶ．日本語のゼロ照応解析は BERT の導入により飛躍的に性能が向上したが [1, 2]，改善の余地が大きく残されている．

大きな課題の一つは学習データの少なさである．アノテーションが付与されている文の数は数万文規模に留まっており，質の高いアノテーションを付与するには言語学的な専門知識が必要なため，大規模にコーパスを拡張することが現実的ではない [3]．

この点に着目し，大規模に利用可能な，日本語と非プロドロップ言語である英語との対訳テキストを利用する研究がなされている [4, 5]．コアとなるアイデアは，日本語におけるゼロ代名詞は英語では明示的に言及されることが多く (図 1(b))，このズレをゼロ代名詞の照応先を解決するための手がかりとできることにある．このアイデアに基づき，Umakoshi らは事前学習とゼロ照応解析の間の中間タスクとして機械翻訳 (MT) を用いることで対訳テキストからゲインが得られることを示した [6]．

同手法は一定の有効性は示されたものの，MT が中間タスクとして適しているかは疑問が残る．様々なタスクの組み合わせを調査した先行研究では，中間タスクとして MT を用いるとスコアが下がる傾向が報告されている [7]．原因として指摘されていることの 1 つは，中間タスクと事前学習との性質の違いが事前学習により得られた知識の忘却を引き起こした可能性である．

そこで本研究では，事前学習に用いられるマスク言語モデル (MLM) により近い翻訳言語モデル (TLM) [8] を中間タスクとする手法を提案する．より乖離の小さいタスクを用いることで効果的に対訳テキストを活用できることを期待する．実験により TLM を用いた場合の方が MT を用いる場合よりも精度が向上することを示す．

## 2　関連研究

事前学習と目的タスクとの間に中間タスクを挟み目的タスクでの精度向上を図る試みがなされてきた [9, 7, 10]．Wang らは MT を中間タスクとした際に自然言語推論や感情分類など様々な目的タスクで精度が悪化したことを報告しており，原因として事前学習と中間タスクとの乖離が破滅的忘却を引き起こした可能性を指摘している [7]．

日本語ゼロ照応解析の精度改善のために多言語テキストを活用することを目的に，比較的大規模な日英対訳コーパスを用いる手法が提案されてきた [4, 5, 6]．従来はパイプライン処理に基づくルー

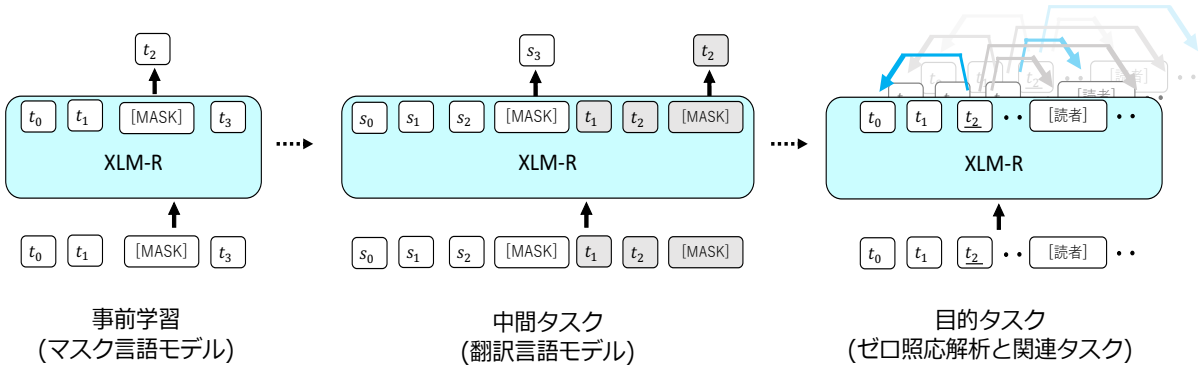


**図 2**　提案手法の概要．**左**: モデルを多言語コーパスを用いてマスク言語モデル (MLM) により事前学習を行う．**中央**: 事前学習したモデルを翻訳言語モデル (TLM) で訓練する．**右**: ゼロ照応解析について訓練する．一部の特殊トークンは図を簡潔にするために省略した．

ルベースの手法が提案されてきたが，誤り伝搬によりその効果は限定的であった [4, 5]．Umakoshi らは MT を中間タスクとした場合にゼロ照応解析の精度を改善できることを示し，特に MLM との同時学習が効果的であることを発見した [6]．MT は事前学習と乖離が大きく中間タスクとして適しているかは疑問であり，本研究ではより乖離の小さい TLM を用いる．

# 3　提案手法

## 3.1　ゼロ照応解析モデル

本研究のゼロ照応解析モデルとして，植田らの *CAModel* [1] を用いる．強力な事前学習済みエンコーダーを考慮し，省略の検出と先行詞の識別は項選択により定式化される (図 2(右))．文書全体を入力し，モデルはその中から述語の各格を埋める項を選択する．外界照応は，照応先に対応する特殊トークンを連結してモデルに入力することにより考慮される．項を持たない場合は，特殊トークン [NULL] を選択する．文書全体を入力とすることによって，自然に文内ゼロ照応だけでなく文間ゼロ照応も扱うことができることに注意されたい．実際には，エンコーダモデルに入力長による制限があり，出来るだけ多くの先行文を含めるように分割した．また，この定式化により，述語と項の間に係り受けの関係がある格解析も同じ枠組みで扱われる．

具体的には，トークン $t_j$ が 述語 $t_i$ の格 $c$ の項である確率は次のように定式化される．

$$P(t_j|t_i, c) = \frac{\exp(s_c(t_j, t_i))}{\sum_{j'} \exp(s_c(t_{j'}, t_i))} \quad (1)$$

$$s_c(t_j, t_i) = \boldsymbol{v}^\top \tanh(W_c \boldsymbol{t}_j + U_c \boldsymbol{t}_i) \quad (2)$$

ここで，$\boldsymbol{t}_i$ はエンコーダにより計算されるコンテキストにおける $t_i$ の埋め込み，$W_c$ と $U_c$ は格ごとの重み行列，$\boldsymbol{v}$ は格の間で共有される重みベクトルである．最も高い確率を持つ $t_j$ を予測とする．全ての述語・格に対しこの計算を行う．

**入力形式**　外界照応などを扱うための特殊トークンを入力系列の末尾に追加する: [著者], [読者], [不特定: 人] を外界照応のために，[NULL] を項を取らないときのために挿入する．また，次のパラグラフで述べる理由で特殊トークン [NA] も挿入する．XLM-R [8] を使う際の慣例に従い，特殊トークン [CLS] 及び [SEP] を系列の最初と最後に挿入する．単語が 2 つ以上のサブワードに分割される場合，最初のサブワードを項選択に用いる．

**マルチタスク学習**　植田ら [1] に従い，一つのモデルを用いて用言の述語項構造解析 (VPA)，体言の述語項構造解析 (NPA)，橋渡し照応解析 (BAR)，共参照解析 (CR) を同時に行う．これら 4 つのタスクは全て同様に式 (1) として定式化することができる．VPA と NPA については格ごとの重みを共有し，BAR，CR にはそれぞれ別の重みを用いる．エンコーダ部分は共有されるために，関連のあるタスクが学習時に互いに影響し合う．

## 3.2　翻訳言語モデル

MT の代わりに翻訳言語モデル (TLM)[8] を用いることが本研究の提案である．この背景にある直感は，TLM は MT よりも MLM に近いためより中間タスクに適している可能性を秘めていることである．

TLM は多言語タスクを目的に対訳テキストを用いた事前学習として提案された [8]．対訳ペアの文章を連結し，その一部をマスクしてモデルに入力

し，マスクされた単語を予測する (図 2(中央))．マスクされた単語と同じ言語のテキストに加えて対言語のテキストからも予測を行うことで，2 言語の間の対応関係をソフトに学習することを期待する．ここでの狙いは，日本語のゼロ代名詞に対応する英単語を予測するような特徴を捉えられるようになることにある．

Conneau らは原言語と目的言語を同じ割合でランダムにマスクしているが，我々の目的においてはマスクする対象を調整することでより効果的に学習が行える可能性がある．1 つの戦略として，単純に英語側の単語のマスクの割合を増やすというものが考えられる．TLM によるゲインの大部分が英語トークンを予測することに起因すると仮定するなら，より多くの英語トークンを予測させることでさらに精度が向上することが期待される．また，日本語のゼロ代名詞に相当する単語の多くは英語では代名詞で現れる点に着目し，英語の代名詞を優先的にマスクする戦略も考えられる．本研究ではこれら 2 つのマスク戦略についての実験も行う．

# 4 実験

中間タスクを TLM にすることの有効性及びマスクする際の設定を変化させることによる影響を検証する．

## 4.1 設定

**モデル**　大規模な多言語コーパス CC-100 で事前学習された大規模事前学習済みモデル XLM-$R_{Large}$ [11] を用いる．

**ゼロ照応解析**　本研究では京都大学ウェブ文書リードコーパス [12] と京都大学テキストコーパス [13] の 2 つのコーパスを用いる．それらのジャンルに基づき，それぞれのコーパスを**ウェブ**，**ニュース**と呼ぶ．これらのコーパスには単語分割，品詞や述語項構造などのアノテーションが付与されている．公開されている設定に基づき，およそ 0.75:0.1:0.15 の割合で訓練，検証，評価用データに分割した．また，訓練の際のハイパーパラメータは付録 C に記載した[1]．訓練時は 2 つのコーパスを混ぜ，評価時はそれぞれ別々に評価を行った．

**機械翻訳・翻訳言語モデル**　読売新聞によって配布されている日英対訳コーパスを用いる[2]．このコーパスはおよそ 130 万文ペアから構成される．ゼロ照応解析は文間の関係を捉える必要があるため，連続する文はゼロ照応解析の際のトークン長 (170) を超えないよう連結される．また，翻訳言語モデルの訓練時，ゼロ照応解析で日本語を入力する位置に英語が含まれないよう，英語トークンが 171 番目の位置から始まるように `[PAD]` トークンを挿入した[3]．分割の割合は学習，検証，評価でそれぞれ 0.9, 0.05, 0.05 とした．訓練の際のハイパーパラメータは付録 D,E の通り．

**代名詞の優先的マスキング**　優先マスク対象とするか否かの判定は spacy [4] を用いて判定した．詳細なルールは付録に示す．対象とする単語は訓練データ中に約 51 万個含まれていた．優先マスク対象と判定された単語に属するサブワードを全てマスクしてもマスクすべきサブワード数に達しない場合はランダムに選んだサブワードをマスクした．

## 4.2 結果

表 1 及び表 2 にウェブ，ニュースにおける結果を示す．XLM-$R_{Large}$ を用いたことで BERT に比べベースラインのスコアが大幅に向上した．**+MT w/ MLM** は MT と MLM の同時学習による手法であり，Umakoshi らの提案手法の中で最も高いスコアを達成している [6]．**+TLM** は TLM を用いるモデル，**+TLM w/ PR masking** は代名詞を優先的にマスクする TLM にそれぞれ対応する．**+MLM** はゲインがデータを追加したことに起因する可能性を検証するために，対訳データの日本語テキストを用いて MLM で訓練したモデルである．

中間タスクによるゲインはデータセットにより傾向が異なる．ウェブにおいては**+MT w/ MLM**, **+TLM w/ PR masking** が最高精度を達成しているが，そのゲインは非常に限定的である．ニュースではどの中間タスクを用いる場合でも精度の向上が見られたが，**+TLM w/ PR masking** がもっとも高い精度を達成した．**+TLM w/ PR masking** はどちらのデータセットでも **+MT w/ MLM** よりも高いか同等の精度を達成しているため，TLM はより中間タスクに適していると考えられる．

また，カテゴリごとの結果に着目すると，ウェブ

1) 訓練の際，シードの設定によっては損失が全く減少しない状態に陥ることがあった．そのようなシードの結果は破棄し，学習を再試行した．

2) https://database.yomiuri.co.jp/about/glossary/

3) 予備実験でこの処理による精度向上が認められたため．

4) https://spacy.io/

**表 1**　**ウェブ**のテストセットにおける精度．太字のスコアは対応するカテゴリでの最高スコアを表す．†: 提案手法．MT w/ MLM, TLM w/ PR masking がもっともよいスコアを達成した．

| 手法 | ウェブ | | | |
|---|---|---|---|---|
| | all | **intra** | **inter** | **exophora** |
| BERT [1] | 70.3 | - | - | - |
| XLM-$R_{Large}$ | 74.7 ±0.607 | 65.7 ±0.586 | 70.1 ±0.580 | 79.6 ±0.828 |
| +MLM | 74.7 ±0.127 | **66.5** ±0.932 | 70.4 ±0.935 | 79.3 ±0.430 |
| +MT w/ MLM | **74.8** ±0.422 | 66.0 ±1.10 | 70.4 ±1.322 | 79.7 ±0.516 |
| +TLM † | 74.7 ±0.558 | 65.1 ±0.477 | **70.6** ±1.85 | **79.8** ±0.448 |
| +TLM w/ PR masking † | **74.8** ±0.377 | 66.4 ±1.34 | 70.3 ±0.508 | 79.6 ±0.587 |

**表 2**　**ニュース**のテストセットにおける精度．太字のスコアは対応するカテゴリでの最高スコアを表す．†: 提案手法．TLM w/ PR masking がもっともよいスコアを達成した．

| 手法 | ニュース | | | |
|---|---|---|---|---|
| | all | **intra** | **inter** | **exophora** |
| BERT [1] | 56.7 | - | - | - |
| XLM-$R_{Large}$ | 61.1 ±0.748 | 65.5 ±0.509 | 55.7 ±1.00 | 60.7 ±1.67 |
| +MLM | 61.4 ±0.528 | 66.4 ±0.553 | 56.1 ±0.853 | 58.9 ±2.07 |
| +MT w/ MLM | 61.7 ±0.570 | 66.1 ±0.834 | 56.3 ±0.901 | 61.1 ±1.01 |
| +TLM † | 61.9 ±0.297 | 66.3 ±0.121 | 56.0 ±0.837 | **61.8** ±1.96 |
| +TLM w/ PR masking † | **62.0** ±0.564 | **66.6** ±0.631 | **56.5** ±0.906 | 61.1 ±1.74 |

のデータセットにおいて**文間**の方が**文内**よりもスコアが高いことは意外な結果である[5]．文間の方が文内より探索空間が広く難しいカテゴリであると考えられてきたが，この結果はその直感に反する．

**英語のマスク割合による影響**　英語を復元するタスクがゼロ照応解析に寄与するのであれば，英語のマスク割合を増加させることでより効果的に学習を行うことができるはずである．この仮説を検証するため，英語のマスク割合を変化させた際のニュースにおけるゼロ照応解析の精度の推移を計測した．また，英語側のマスク復元の精度も同時に計算し，これら 2 つの関係を図 3 に示す．

結果から，TLM の精度がマスクの割合が増えるほど一貫して低下していることがわかるが，ゼロ照応解析とのはっきりした関連は見られなかった．これら二つの相関係数を計算したところ 0.567 であり，弱いながらも相関があることがわかった．

## 5　おわりに

本研究では，事前学習と中間タスクとの乖離に着目し，より事前学習に近い TLM を中間タスクとする手法を提案した．実験により，TLM が MT よりもよい中間タスクであることを示した．今後は，対訳テキスト以外のリソースの活用についても考えていきたい．

5）ニュースでこの傾向が見られないのは KC では長い文書を分割して入力しており，照応先がテキスト中に現れないような事例が多く含まれているためと考えられる．

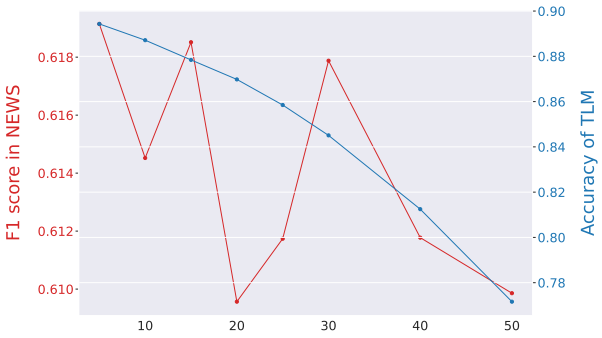


**図 3**　英語側のマスク割合を変化させた際のゼロ照応解析及び TLM の精度の変化

## 謝辞

対訳コーパスを提供していただいた読売新聞東京本社に深く感謝いたします．

## 参考文献

[1] Nobuhiro Ueda, Daisuke Kawahara, and Sadao Kurohashi. BERT-based cohesion analysis of Japanese texts. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 1323–1333, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[2] Ryuto Konno, Yuichiroh Matsubayashi, Shun Kiyono, Hiroki Ouchi, Ryo Takahashi, and Kentaro Inui. An empirical study of contextual data augmentation for Japanese zero anaphora resolution. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 4956–4968, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[3] 阿部航平, 河原大輔, 黒橋禎夫. クラウドソーシングを用いた日本語述語項構造タグ付きコーパスの拡張. 言語処理学会 第 26 回年次大会, pp. 1547–1550, 茨城, 2020.

[4] Hiromi Nakaiwa. Automatic extraction of rules for anaphora resolution of Japanese zero pronouns in Japanese-English machine translation from aligned sentence pairs. **Machine Translation**, Vol. 14, No. 14, pp. 247–279, 1999.

[5] 古川智雅, 中澤敏明, 柴田知秀, 河原大輔, 黒橋禎夫. 対訳コーパスを用いたゼロ照応タグ付きコーパスの自動構築. 言語処理学会 第 23 回年次大会, pp. 382–385, つくば, 2017.

[6] Masato Umakoshi, Yugo Murawaki, and Sadao Kurohashi. Japanese zero anaphora resolution can benefit from parallel texts through neural transfer learning. In **Findings of the Association for Computational Linguistics: EMNLP 2021**, pp. 1920–1934, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[7] Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, and Samuel R. Bowman. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 4465–4476, Florence, Italy, July 2019. Association for Computational Linguistics.

[8] Alexis CONNEAU and Guillaume Lample. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 32. Curran Associates, Inc., 2019.

[9] Jason Phang, Thibault Févry, and Samuel R. Bowman. Sentence encoders on STILTs: Supplementary training on intermediate labeled-data tasks. **arXiv:1811.01088**, 2018.

[10] Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. Intermediate-task transfer learning with pretrained language models: When and why does it work? In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 5231–5247, Online, July 2020. Association for Computational Linguistics.

[11] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics.

[12] Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. Building a diverse document leads corpus annotated with semantic relations. In **Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation**, pp. 535–544, Bali, Indonesia, November 2012. Faculty of Computer Science, Universitas Indonesia.

[13] Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hasida. Construction of a Japanese relevance-tagged corpus. In **Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)**, Las Palmas, Canary Islands - Spain, May 2002. European Language Resources Association (ELRA).

## A　優先的マスキングのルール

優先マスク対象の単語は spacy (v. 3.2.1) によって付与される POS タグに基づき決定した．具体的には，

- tag_ 属性の値が PRP
- pos_ 属性の値が PRON かつ tag_ 属性の値が DT

のいずれかを満たす単語を優先的なマスク対象とした．

## B　事前学習とファインチューニングのギャップ

コーパスは基本句単位でアノテーションが付与されている一方，事前学習モデルは単語分割なしで訓練されており，ギャップがある．出来るだけ事前学習のときの条件に近づけるため，トークナイズの際は単語分割された後の単語に対しトークナイズを行い，文頭に現れる特殊トークンを削除するという前処理を行なった．

## C　ゼロ照応解析におけるハイパーパラメータ

| 項目 | 値 |
|---|---|
| オプティマイザ | AdamW |
| Adam の eps | $1 \times 10^{-8}$ |
| 重み減衰 | 0.01 |
| エポック数 | 4 |
| バッチサイズ | 8 |
| 学習率 | $5.0 \times 10^{-5}$ |
| ウォームアップの比率 | 0.1 |
| 損失関数 | Cross entropy |
| ドロップアウト (BERT layer) | 0.1 |
| ドロップアウト (output layer) | 0.0 |
| 学習率スケジューラ | linear_schedule-_with_warmup [6] |

**表 3**　ゼロ照応解析におけるハイパーパラメータ

## D　機械翻訳におけるハイパーパラメータ

| 項目 | 値 |
|---|---|
| オプティマイザ | Adam |
| Adam のパラメタ | $\beta_1$=0.9, $\beta_2$ = 0.98 |
| Adam の eps | $1 \times 10^{-6}$ |
| 重み減衰 | 0.01 |
| エポック数 | 50 |
| バッチサイズ | Approx. 500 |
| 学習率 | $5.0 \times 10^{-6}$ |
| ウォームアップのエポック数 | 5 |
| 損失関数 | Lable-smoothed cross entropy |
| 平滑化定数 | 0.1 |
| ドロップアウト (BERT & Dec.) | 0.1 |
| 学習率スケジューラ | polynomial decay |

**表 4**　機械翻訳におけるハイパーパラメータ

## E　翻訳言語モデルのハイパーパラメータ

| パラメータ | 値 |
|---|---|
| オプティマイザ | AdamW |
| Adam のパラメタ | $\beta_1$=0.9, $\beta_2$ = 0.999 |
| オプティマイザのイプシロン | $1 \times 10^{-8}$ |
| 重み減衰 | 0.01 |
| エポック数 | 20 |
| バッチサイズ | 128 |
| 学習率 | $5.0 \times 10^{-6}$ |
| ウォームアップ | 2 エポック |
| 損失関数 | cross entropy |
| ドロップアウト | 0.1 |
| 学習率スケジューラ | linear_schedule-_with_warmup [6] |

**表 5**　翻訳言語モデルのハイパーパラメータ

6)　https://github.com/huggingface/transformers/blob/v2.10.0/src/transformers/optimization.py#L47