非自己回帰的な生成と事前学習を用いた機械翻訳への試み

中村 朝陽 † 鶴岡 慶雅 ‡
† 東京大学 工学部電子情報工学科
‡ 東京大学 大学院情報理工学系研究科
{nakam, tsuruoka}@logos.t.u-tokyo.ac.jp

1 はじめに

国際化の進む現代社会には様々な言語のテキストデータが溢れており、機械翻訳のニーズが増え続けている。更にリアルタイム性を要する対話やSNSにおいては、高速に翻訳を行う必要性も高まってきている。

Vaswani ら [1] の提案した自己注意機構を用いたTransformerが登場し、ニューラルネットワークを用いた機械翻訳に大きな影響を与えた。一方でTransformerにおける自己回帰的 (autoregressive) な復号化に対する課題も指摘されている。自己回帰的な復号化では出力系列を先頭から1トークンずつ生成し、モデル自身が出力したトークンを再度モデルに入力することによって系列データを出力しているが、復号にかかる時間が系列長に比例して長くなってしまう。また各ステップで最も確率の高いトークンを貪欲に選択しても局所解しか得られず、最尤の出力系列全体を予測することが難しい。この問題を緩和するために複数の候補を加味するビームサーチが広く使われており精度の向上に寄与しているものの、計算時間の大きなオーバーヘッドとなることが知られている。

そこで Gu ら [2] によって非自己回帰的 (non-autoregressive) な復号化が提唱され、Ghazvininejad ら [3] 等がよりシンプルかつ高性能で、計算量が出力系列長に依存せず定数回のステップで実行できる復号化手法を提案した。これらの非自己回帰的なアプローチでは計算量を大きく減らすことが出来る代わりに、出力トークン間の依存関係を捉えにくいため自己回帰的な手法と比べ性能が低下することが知られている。

そこで本研究では計算コストの低い非自己回帰的な復号化を用い、加えて Devlin ら [4] の BERT に代表されるような事前学習を取り入れることで高品質な翻訳を行うことを目的とする。

2 関連研究

2.1 非自己回帰的な復号化

Gu ら [2] は自己回帰的な復号化の問題点と、単純な非自己回帰的な実装では上手く翻訳を行うことが出来ないという課題を指摘した上で、Fertility という概念を導入することを通じて初めて非自己回帰的な機械翻訳を行う手法を提案した。

また Ghazvininejad ら [3] は符号化されたソース文に条件づけられたマスク付き言語モデリングを行い、マスクされた単語の予測を一定回数繰り返すような復号化を提案した。これは Gu ら [2] のような純粋な非自己回帰的な手法とは異なり、出力系列長には依存しないものの 4〜10 回程度の一定の繰り返しによる復号化を行っている。ここで確信度の高いトークンから順に確定していくことで、一部の出力トークン間の依存関係をモデル化している。

また 2019 年には Gu ら [5] は、挿入と削除という演算を Transformer 上で行うモデルを提案した。これらのモデルも一定回数の復号化により高速かつ精度の高いテキスト生成を可能にした。

2.2 事前学習を用いた自然言語処理

事前学習を用いて自然言語理解タスクを大きく進めた Devlin らの BERT ベースアーキテクチャを Seq2Seq タスクに応用する研究も多く行われている。Dong ら [6] の提案した UniLM では、事前学習を用いて BERT に引けを取らない自然言語理解スコアを出しつつ、同一の事前学習済みモデルを用いて要約や質問応答などのタスクで高い性能を示した。また Song ら [7] は Vaswani らの Transformer のアーキテクチャを元に事前学習を行った上で Seq2Seq タスクを試みた。事前学習により単一言語コーパスから多くの情報を獲得したことにより、教師なし英仏翻訳で高い性能を示した。

事前学習を Seq2Seq タスクに応用するためのアーキテクチャとして従来は符号化器のみからなる BERT ベースのアーキテクチャが採用されてきたが、Raffel [8] らは入出力の形式を工夫することで符号化器と復号化器からなる Transformer アーキテクチャでもマスク付き言語モデリングのような事前学習の恩恵を同じように受けられることを示した。これにより翻訳や要約などの Seq2Seq タスクに対してもより自然な形で BERT で行われたような事前学習を適用できるようになった。

3 提案手法

本研究では大規模な単一言語コーパスを用いて事前学習を行い、その後対訳コーパスなどのデータを用いて翻訳などの Seq2Seq タスクを非自己回帰的なアプローチで解く手法を提案する。Raffel らのように Seq2Seq タスクに事前学習を活用することにより、限られた対訳コーパスのみならず大規模な単一言語コーパスをモデルが学習することが出来るため、精度の向上が期待される。また、Ghazvininejad らのように非自己回帰的なアプローチを適用することにより、実行時間が出力テキストの系列長に比例することなく一定時間で計算することが出来るようになるため、高速な演算が可能になる。

3.1 アーキテクチャ

本研究では Vaswani らの Transformer アーキテクチャを用いる。6 層の Transformer レイヤーからなる符号化器と復号化器で、符号化器には自己注意機構、復号化器には自己注意機構に加え、ソース-ターゲット間の注意機構が含まれる。Ghazvininejad らの復号化手法を用いるため、復号化器の自己注意機構において将来のトークンを見ないようにするための左から右への注意へのマスクは用いていない点が通常の Transformer とは異なる。ハイパーパラメータに関しては Transformer-Base と同様で、隠れ層の次元は 512、フィードフォワード層の次元は 2048、ヘッド数は 8 である。

3.2 事前学習

多言語を扱う翻訳タスクの事前学習として、ソース言語とターゲット言語のそれぞれで事前学習を試みた。本提案手法では符号化器と復号化器からなるアーキテクチャを用いているため、BERT での事前学習をそのまま行うことは難しい。Raffel らは符号化器と復号化器からなるアーキテクチャでマスク言語モデリングを行っているが、マスクされたトークンを自己回帰的に出力している。そこで本研究では非自己回帰的な生成を行うために、マスクされた単語全てを 1 ステップで並列的に予測する条件付きのマスク言語モデリングを事前学習タスクとして提案する。概要を図 1 に示す。連続する 2 文を対象とし、1 文目を符号化器に、マスクをかけた 2 文目を復号化器に入力し、2 文目のマスクされた単語を当てるように学習する。交差エントロピー誤差を目的関数としてパラメータを更新する。また正則化としてラベルスムージング [9] ($e_{ls} = 0.1$) を適用した。

3.3 翻訳タスクの学習

翻訳タスクの学習と、評価時の推論方法については Ghazvininejad らの手法を用いた。

翻訳タスクの学習は事前学習とほぼ同じで、ソース言語の文を符号化器に、マスクされたターゲット言語の文を復号化器に入力する。そしてターゲット文のマスクされたトークンを予測し、交差エントロピー誤差を最小化する。図 2 に概要を示す。またマスクされたトークンを予測するという非自己回帰的な生成手法では、文長を変化させることができない。そのためここで符号化器の入力から出力トークン長を予測するタスクを追加する。BERT の CLS トークンのように LENGTH トークンを符号化器に入力し、そのトークンに対応する符号化器の出力と実際の出力文長との誤差を交差エントロピー誤差に追加する。

3.4 翻訳タスクの推論

推論時にはまず入力文を符号化器に入力し、出力系列の長さを予測する。1 バッチ内の予測長がそれぞれ異なると並列にミニバッチ処理を行うことが難しくなるため、バッチ内の予測を集め上位 l 件の出力長候補を決める。そうしたら出力系列長候補を一つ選び、その個数分のマスクトークンを復号化器に入力し、各マスクのトークンを予測する。そして予測スコアの高いものから幾つかを選択し、マスクの予測結果として確定させる。その後は確定した予測結果とそれ以外のマスクトークンを繰り返し復号化器に入力する。本研

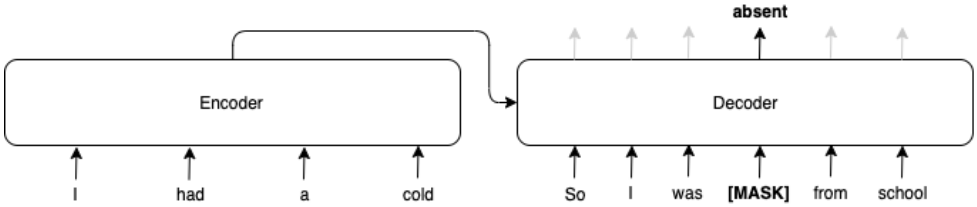

図 1: 事前学習のための条件付きマスク言語モデリング

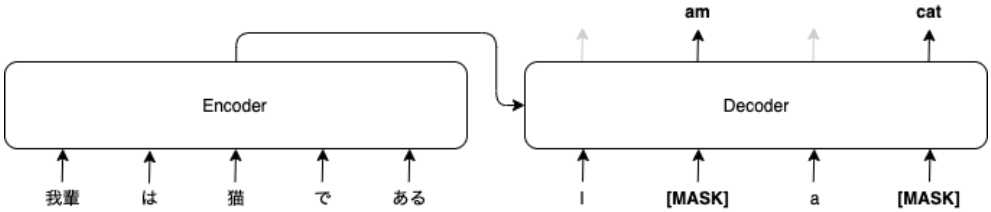

図 2: 非自己回帰的に翻訳文を生成するための学習

表 1: 推論時の復号化の流れ

ソース文	吾輩は猫である
出力長予測	[MASK] [MASK] [MASK] [MASK]
$t = 1$	I [MASK] [MASK] cat
$t = 2$	I am a cat

究では Ghazvininejad らに倣って、イテレーション数 T、出力系列長 N のとき、イテレーション t ではマスクの予測スコア上位 Nt/T 件が確定するようにした。このアルゴリズムの例を表 1 に示す。

4 実験

4.1 実験設定

事前学習には英独の Wikipedia を利用した。英語 Wikipedia と独語 Wikipedia ではコーパスの大きさが異なるため、共に 1,000 万文ずつ利用した。また翻訳データセットとしては WMT'14 EN-DE を用いた。トークン化には Sennrich ら [10] の subword-nmt[1]を利用して語彙サイズ 30,000 で分割した。また予備実験で事前学習の後に翻訳タスクを学習させた場合には精度の向上が見られなかったため、翻訳タスクを一度学習したモデルに対して事前学習を行い、その後再び翻訳タスクを学習させた。

ベースとなる通常の自己回帰モデルは Transformer-Base モデルでビーム幅が 5 である。非自己回帰モデルでは 3.3 - 3.4 節にあるような訓練と推論を行った。非自己回帰モデルでは共に出力のイテレーション数 T は 10、出力系列長の候補数 l は 2 とした。

4.2 結果

テストデータに対する翻訳結果の例を表 2 に、BLEU [11] スコアを表 3 に示す。

非自己回帰のみのモデルから事前学習を取り入れることで精度がわずかに向上した。また自己回帰モデルに比べ非自己回帰モデル 2 つは 1.3 倍程度の速度で推論を行うことができた。

表 2 の 1 つ目の出力例ではよく翻訳できていることが確認できる。参照文 1 と出力文 1 の唯一の違いは “shut” と “close” のみで、事前学習によりこれら 2 単語が近い意味であることを学習できたためであると推測できる。一方で 2 つ目の出力例では似たような単語を用いているものの誤った意味の文を出力してしまっている。参照文と出力文での系列長が異なることから、系列長予測に失敗してしまった影響で文意を誤ってしまったことなどが理由として考えられる。

[1]https://github.com/rsennrich/subword-nmt

表 2: 翻訳結果の例

入力文 1	Ich kann nur meine Augen schließen und sie langsam wieder öffnen ...
参照文 1	I can only shut my eyes and slowly open them again ...
出力文 1	**I can only close my eyes and open them slowly again ...**
入力文 2	In der Museumswerkstatt wird ebenfalls gewerkelt .
参照文 2	Activities will also be taking place in the museum workshop .
出力文 2	**The Museum Workshop is also organised .**

表 3: 翻訳スコア (BLEU)

モデル	EN-DE	DE-EN	速度
自己回帰 (Transformer)	27.3	31.1	1.0x
非自己回帰	23.5	27.1	1.3x
非自己回帰 + 事前学習	**24.6**	**27.5**	**1.3x**

4.3 考察

表 3 のように事前学習を取り入れることで翻訳精度が向上したが、自己回帰モデルとの差は依然大きいままであり事前学習はあまり大きな影響を与えなかった。これは事前学習では単一言語のタスクを解いていた一方で、翻訳タスクでは複数言語のタスクを学習しており、事前学習タスクと翻訳タスクが大きく異なっていたことが原因として考えられる。また事前学習用データセットのドメインと翻訳データセットのドメインが異なることや、単純にモデルの初期値として事前学習済みモデルを使う以外の事前学習の活用が必要である可能性も考えられる。

5 おわりに

本研究では高速だが精度が落ちてしまう非自己回帰的な復号化手法に対し、事前学習を用いることで精度の減少を抑えた。しかし事前学習だけでは非自己回帰モデルと自己回帰モデルの差は依然として大きい。多言語での事前学習手法の再考や、モデルの蒸留、逆翻訳 [12]、異なるアーキテクチャのモデルによる改善などが期待される。

参考文献

[1] Vaswani, et al. Attention is all you need. In *NeurIPS*. 2017.

[2] Gu, et al. Non-autoregressive neural machine translation. In *ICLR*, 2018.

[3] Ghazvininejad, et al. Mask-predict: Parallel decoding of conditional masked language models. In *EMNLP*, 2019.

[4] Devlin, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, 2018.

[5] Gu, et al. Levenshtein transformer. In *NeurIPS*. 2019.

[6] Dong, et al. Unified language model pre-training for natural language understanding and generation. In *NeurIPS*, 2019.

[7] Song, et al. Mass: Masked sequence to sequence pre-training for language generation. In *ICML*, 2019.

[8] Raffel, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*, 2019.

[9] Szegedy, et al. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[10] Sennrich, et al. Neural machine translation of rare words with subword units. In *ACL*, 2016.

[11] Papineni, et al. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[12] Edunov, et al. Understanding back-translation at scale. In *ACL*, 2018.