Dualformer の双方向性を活かしたパラメータ共有による機械翻訳

川村吏輝　秋葉友良
豊橋技術科学大学
{kawamura.riki.zc, akiba.tomoyoshi.tk}@tut.jp

塚田元
愛知産業大学
tsukada@asu.ac.jp

概要

近年, ニューラル機械翻訳の分野では Transformer が主流であるが, 翻訳問題の双方向性を単一モデルで解く手法として Dualformer が提案されている. 本研究では, Dualformer の双方向性をさらに活かした性能向上とパラメータ効率化を目的とし, 同一モデルアンサンブルとデコーダのパラメータ共有を提案する. 同一モデルアンサンブルでは, Dualformer の単一モデル内で異なる経路を用いたアンサンブルを実現する. また, パラメータ共有では 2 つのデコーダを完全に共有することで, パラメータの削減を行う. 実験では日英間での翻訳タスクにおいて, 提案手法は統計的に有意な精度の向上を達成し, かつ大幅な軽量化を実現しても性能が劣化しないことを示した.

1　はじめに

近年, ニューラル機械翻訳 (NMT) の分野では, Transformer [1] をベースとしたモデルが目覚ましい成果を上げている. 翻訳というタスクには, ある言語から別の言語への翻訳（例：日英）と, その逆方向の翻訳（例：英日）という対になる問題が必ず存在し, この性質は双方向性 (Duality) と呼ばれる. この双方向性を利用して, 互いの翻訳精度を高め合う枠組みとして双方向学習 (Dual Learning) が知られている.

Xia ら [2] は, この双方向性の問題をモデルレベルで解決する手法 (Model-level dual learning) を提案した. 彼らが提案した手法は, Transformer の構造を応用し, 単一のモデルで双方向の翻訳を可能にするものである. Dualformer は 2 つのデコーダを持ち, 一方を疑似的なエンコーダとして振る舞わせることで, パラメータを共有しながら双方向の学習を行う. 森下ら [3] や Chien[4] は Xia らと同様のアーキテクチャを持つ Transformer ベースのモデルの学習に補助タスクを導入し, モデル性能の改善を達成している. また, 加藤ら [5] は同様のアーキテクチャのモデルがドメインシフトに頑健なモデルであることを示している. 本稿では, Chien らや加藤らに従い, 拡張された Transformer を Dualformer と呼ぶ.

従来の Dualformer には, さらなる性能向上とパラメータ効率化の余地が残されている. Dualformer はエンコーダとデコーダの役割をスイッチさせることでパラメータ共有を行っているが, 2 つのデコーダ自体は独立して存在しているためである.

本研究では, Dualformer の双方向性をさらに活かした 2 つの手法を提案する. 第一に, 同一モデルアンサンブルである. 各翻訳方向を多言語モデルのタスクとして捉え, 学習データを混合することで, すべてのデコーダが双方向の翻訳能力を獲得するように学習を行う. これにより, 単一のモデル内で異なる経路を用いたアンサンブルが可能となり, 翻訳精度の向上が期待できる. 第二に, デコーダの完全共有 (Shared Dualformer) である. 従来 2 つ存在したデコーダを 1 つに統合し, タスク実行時に複製して利用することで, Dualformer としての動作を維持しつつ, 大幅なパラメータ削減と効率化を図る.

本稿では, 提案手法の有効性を検証するために, 日英翻訳タスク（KFTT および ASPEC）を用いた評価実験を行う. 実験の結果, 提案手法である同一モデルアンサンブルによる精度の向上, およびデコーダ共有によるパラメータ数の削減と性能維持が可能であることを示す.

2　Dualformer

Dualformer[2] は, 機械翻訳における双方向性 (Duality) をモデル構造レベルで解決するために提案された手法である. 従来の Transformer がエンコーダとデコーダのペアで構成され, 一方向の翻訳（例：言語 A から言語 B）のみを学習するのに対し, Dualformer は単一のモデルで双方向（A → B および

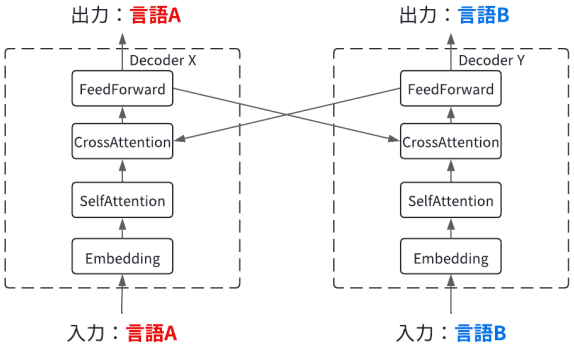

図 1　Dualformer の概要図

B → A）の翻訳を行うことが可能である．

2.1　モデル構造

Dualformer は，2 つのデコーダ（本稿では Decoder X, Decoder Y とする）によって構成される．通常の Transformer に見られる独立したエンコーダは保持しておらず，翻訳方向に応じて一方のデコーダが疑似的なエンコーダ (Pseudo-Encoder) として振る舞うことで，エンコーダ・デコーダ構造を動的に形成する．

2.2　翻訳プロセス

言語 A から言語 B への翻訳 (A → B) において，Dualformer は以下のように動作する．まず，Decoder X がエンコーダの役割を担う．Decoder X 内の Cross-Attention 層を無視（マスク）し，Self-Attention 層と Feed-Forward 層のみを使用することで，入力された言語 A の系列特徴量を抽出する．この処理は，通常の Transformer におけるエンコーダの動作と等価である．続いて，Decoder Y がデコーダとして動作する．Decoder Y は，通常のデコーダと同様に Cross-Attention 層を用いて，疑似エンコーダである Decoder X の出力を参照しながら，目的言語である言語 B の系列を生成する．

3　提案手法

本研究では，Dualformer の双方向性をさらに活用し，翻訳精度の向上とモデルの軽量化を実現する 2 つの手法を提案する．

3.1　同一モデルアンサンブル

従来の Dualformer では，言語 A から言語 B への翻訳 (A → B) において，Decoder X をエンコーダ，Decoder Y をデコーダとして使用する経路が固定されていた．多言語モデルの学習フレームワークを Dualformer に適用することで，デコーダ方向 X → Y と Y → X の両方で同じ言語方向の翻訳が可能になる．DecoderX のエンコーダ + DecoderY のデコーダと DecoderY のエンコーダ + DecoderX のデコーダのそれぞれの出力を平均化し，同一モデルアンサンブルを行う．

3.2　Shared Dualformer

Dualformer は 2 つのデコーダ (X, Y) を有しているが，これらは構造的に同一である．そこで，さらなるパラメータ効率化のために，これらのデコーダのパラメータを完全に共有する手法 (Shared Dualformer) を提案する．

本手法では，単一の「Decoder S」のみを保持する．学習や翻訳のタスク実行時には，この Decoder S を複製 (コピー) して配置することで，仮想的に 2 つのデコーダが存在するかのように振る舞わせる．すなわち，従来の Dualformer における Decoder X と Decoder Y の役割を，どちらも同一のパラメータを持つ Decoder S が担うことになる．

これにより，Dualformer としての動作メカニズムを維持したまま，パラメータ数を大幅に削減することが可能となる．実験の章で後述するが，この共有化を行っても十分な層の深さを確保すれば，翻訳性能を劣化させることなくモデルの軽量化が実現できる．

4　実験設定

本節では，提案手法の有効性を検証するために行った実験の条件について述べる．

4.1　データセットと前処理

実験には，日英翻訳のベンチマークとして広く用いられている 2 つのコーパスを使用した．

表 1　実験に使用したデータセットの統計．ASPEC はパラメータ共有の実験のみで使用された．

言語対	データセット	訓練	開発	テスト
Ja ↔ En	KFTT	440k	1,166	1,160
	ASPEC	1M	1,790	1,812

1 つ目は京都フリー翻訳タスク (KFTT) である．KFTT は京都に関連する Wikipedia 記事の日英対訳データであり，学習データ数は約 44 万文である．2 つ目はアジア学術論文抜粋コーパス (ASPEC) であ

表 2 同一モデルアンサンブルの実験結果 (KFTT) *: Dualformer_base に対して統計的に有意 ($p < 0.05$)

Model	en → ja		ja → en	
	BLEU	COMET	BLEU	COMET
Transformer_base	21.82	0.8260	22.21	0.7530
Transformer_multi	22.13	0.8262	21.98	0.7497
Dualformer_base	22.61	0.8328	23.70	0.7601
Dualformer_multi ($X \rightarrow Y$)	22.33	0.8265	22.46	0.7534
Dualformer_multi ($Y \rightarrow X$)	21.90	0.8286	22.67	0.7505
Dualformer_multi_ensemble	**23.43**	**0.8394***	**24.80**	**0.7599**

表 3 KFTT および ASPEC におけるデコーダパラメータ共有の実験結果 *: Transformer に対して統計的に有意 ($p < 0.05$)

Dataset	Model	Params	en → ja		ja → en	
			BLEU	COMET	BLEU	COMET
KFTT	Transformer	~154M	21.82	0.8260	22.21	0.7530
	Dualformer	94M	22.61	0.8328*	23.70	0.7601*
	Shared Dualformer	**66M**	**22.80**	**0.8314***	**23.58**	**0.7587***
ASPEC	Transformer	~140M	39.57	0.8893	28.43	0.7995
	Dualformer	85M	39.91	0.8913*	28.99	0.8038*
	Shared Dualformer	**58M**	**39.78**	**0.8898**	**28.81**	**0.8026**

る. ASPEC は科学技術論文の日英対訳データであり, 本実験では学習データとして最初の 100 万文を使用した. ASPEC は主にパラメータ共有手法 (Shared Dualformer) の検証に用いた.

トークン化には SentencePiece[6] を使用し, サブワード単位での分割を行った. また, 提案手法 1 (同一モデルアンサンブル) の評価においては, 多言語モデル化のために, 日英 (ja2en) と英日 (en2ja) の対訳データを混合して学習データを構築した. これにより, 言語方向を区別せず 1 つのタスクとして学習させた.

4.2 モデル設定

モデルの実装と学習には fairseq[7] を用いた. 各モデルのハイパーパラメータは Vaswani ら [1] の Base Model と同じ設定とした. 提案手法 1 (同一モデルアンサンブル) の検証では, 通常の Transformer および Dualformer, それぞれ多言語化したモデルを比較した. 提案手法 2 (デコーダ共有) の検証では, 通常の Transformer, 通常の Dualformer, そして提案手法である Shared Dualformer を比較した.

4.3 評価手法

モデルの評価には, 開発セットに対する損失が最小となるチェックポイントを使用した. 翻訳精度の評価指標として, 以下の 2 つを用いた. BLEU : 出力文と参照文 (正解) の単語の一致度に基づく指標. COMET[8]: 出力文と参照文の意味的な類似度をニューラルネットワークを用いて判定する指標. また, 提案手法とベースラインとの間の性能差が統計的に有意であるかを確認するために, ブートストラップ法に基づく対応のある t 検定を行った.

5 実験結果

本節では, 提案手法である同一モデルアンサンブルおよびデコーダのパラメータ共有の評価結果について述べる.

5.1 同一モデルアンサンブルの結果

表 1 に, KFTT データセットを用いた同一モデルアンサンブルの実験結果を表 2 に示す. 比較対象として, 通常の Transformer (Transformer_base), 多言語モデル化した Transformer (Transformer_multi), 通常の Dualformer (Dualformer_base), および提案手法である多言語 Dualformer (Dualformer_multi) のスコアを記載している.

実験の結果, 提案手法である Dualformer_multi_ensemble は, 英日翻訳 (en2ja) において BLEU スコア 23.43, COMET スコア 0.8394 を記録し, 朝日翻訳 (ja2en) においても BLEU スコア 24.80, COMET スコア 0.7599 を記録した. これらは, 比較手法の中で最も高い数値であり, 特に Dualformer_base と比較して統計的に有意 (p<0.05) な性能向上が確認された.

Dualformer_multi の単体出力（x → y または y → x の片方向のみ使用）では, ベースラインと同程度の性能にとどまっていることから, 単一モデル内で異なる経路（エンコーダ役とデコーダ役の入れ替え）による出力を統合するアンサンブル効果が, 精度の向上に大きく寄与していることがわかる.

5.2 デコーダ共有の結果

5.2.1 翻訳精度とパラメータ効率

次に, デコーダを完全に共有した Shared Dualformer の性能検証を行った. 表 3 に, KFTT と ASPEC データセットにおける結果を示す.

KFTT における実験では, Shared Dualformer は英日翻訳で通常の Dualformer と同等, あるいは Transformer (BLEU 21.82, COMET 0.8260) よりも高い性能を示した. 特筆すべきはパラメータ数である. 通常の Dualformer が 94M, Transformer (2 モデル合計) が約 154M であるのに対し, Shared Dualformer は 66M まで削減されている. これは, デコーダを共有することで約 30%のパラメータ削減を実現しながらも, 翻訳性能を維持・向上できることを示している.

ASPEC における実験でも同様の傾向が見られた. Shared Dualformer (58M) は, 英日翻訳において BLEU 39.78 を記録し, パラメータ数がより多い Transformer (140M, BLEU 39.57) を上回る結果となった.

5.2.2 層の深さと精度の関係

Shared Dualformer における層の深さが翻訳精度に与える影響について分析する. 図 2 は, モデルのパラメータ数と COMET スコアの関係を層数（Layer 1, 3, 6）ごとにプロットしたものである.

実験の結果, 層数が浅い場合（Layer 1, 3）, Shared Dualformer は Transformer や通常の Dualformer と比較して性能の劣化が見られ, 有意に劣るケースが確認された. しかし, 層を十分に深く（Layer 6）設定した場合は, パラメータを共有しても性能劣化は見られず, むしろ Transformer に対して有意に優れる結果が得られた. これは, デコーダを共有することでモデルの表現力が制約されるものの, 層を深くすることでその容量不足を補い, 共有による正則化効果あるいは学習効率の向上が上回ったためと考えられる. したがって, Shared Dualformer の実用には十分な層の深さを確保することが重要である.

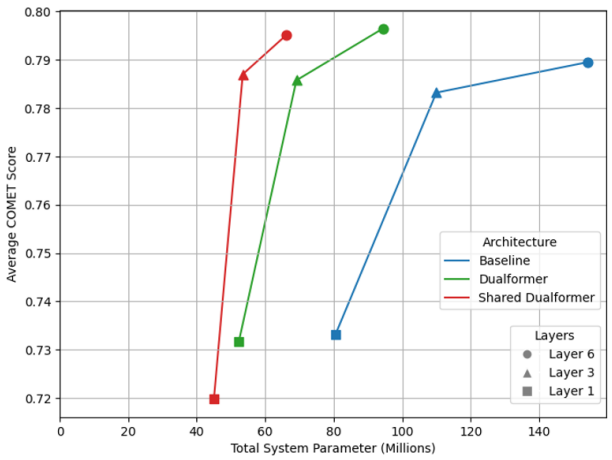

図 2 KFTT におけるレイヤー層の変化 (縦軸：双方向の平均 COMET スコア, 横軸：モデルの総パラメータ数)

6 おわりに

本研究では, 機械翻訳における双方向性をモデル構造レベルで活用する Dualformer に着目し, さらなる翻訳精度の向上とパラメータ効率化を実現する 2 つの手法を提案した.

1 つ目の提案手法である同一モデルアンサンブルでは, 多言語モデル化によって単一モデル内に複数の翻訳経路を構築した. 実験の結果, これらを統合することで, ベースラインおよび通常の Dualformer と比較して, 特に英日翻訳において統計的に有意な性能向上を達成した.

2 つ目の提案手法であるデコーダのパラメータ共有（Shared Dualformer）では, 従来独立していた 2 つのデコーダを完全に共有することで, 従来の Dualformer と比較して約 30%のパラメータ削減を実現した. 実験では, 十分な層の深さを確保することで, 大幅な軽量化を行っても翻訳性能が劣化せず, むしろベースラインを上回る結果が得られることを確認した. これは, 計算資源が限られた環境下での高精度な双方向翻訳モデルの構築に寄与する成果である.

今後の展望として, 本研究では日英・英日という言語間距離の遠い言語対で検証を行ったが, 英語とドイツ語のような言語間距離が近い言語対での有効性の検証が挙げられる. また, 低資源言語ペアにおける性能検証や, 3 言語以上を扱う多言語モデルへの拡張についても取り組んでいきたい.

謝辞

本研究は JSPS 科研費 23K11118 の助成を受けたものです.

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[2] Yingce Xia, Xu Tan, Fei Tian, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Model-level dual learning. In Jennifer Dy and Andreas Krause, editors, **Proceedings of the 35th International Conference on Machine Learning**, Vol. 80 of **Proceedings of Machine Learning Research**, pp. 5383–5392. PMLR, 10–15 Jul 2018.

[3] 森下睦, 鈴木潤, 永田昌明. 双方向学習と再現学習を統合したニューラル機械翻訳. 言語処理学会 第 25 回年次大会, 2019.

[4] Jen-Tzung Chien and Wei-Hsiang Chang. Dualformer: a unified bidirectional sequence-to-sequence learning. In **ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 7718–7722. IEEE, 2021.

[5] 加藤龍兵, 秋葉友良, 塚田元. 双対学習機械翻訳モデルのドメインシフトに対する頑健性の検証. 言語処理学会 第 30 回年次大会, 2024.

[6] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. **arXiv preprint arXiv:1808.06226**, 2018.

[7] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In Waleed Ammar, Annie Louis, and Nasrin Mostafazadeh, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)**, pp. 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[8] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics.