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

加藤龍兵　秋葉友良　塚田元
豊橋技術科学大学
{kato.ryuhei.vb,akiba.tomoyoshi.tk,tsukada.hajime.hl}@tut.jp

概要

近年，Transformer ベースモデルがニューラル機械翻訳（NMT）の分野で盛んに研究されている．一般に Transformer は単方向翻訳を行うように学習されるが，翻訳問題には逆方向の翻訳問題が必ず存在する．単一モデルで双方向の翻訳を行うモデルには Dualformer がある．Dualformer は両方向の Transformer ベースモデルのデコーダ部を組み合わせたモデルである．実験では，ドメインの内外問わず独英および日英の両言語対で Dualformer が Transformer の性能を超える結果となった．この結果は，Dualformer が言語対にかかわらず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．

1　はじめに

近年，ニューラル機械翻訳（NMT）の分野は発展が著しい．特に，Transformer [1] に基づく NMT モデルは，従来の Recurrent Neural Network に基づく NMT モデルの性能を多くのタスクで凌駕している．

一般的な Transformer ベース NMT モデルはある言語対について単方向の翻訳を行うように学習される．一方，日英翻訳に対する英日翻訳のように翻訳問題には必ず逆向きの問題が存在し，これらは双対問題と呼ばれる．双対問題は他の系列変換タスクにも見られ，例えば音声認識と音声合成は言語音声・書き起こしテキスト間の双対問題である．

双対問題を解くことで双方のモデル性能を高める手法は双対学習（Dual Learning）と呼ばれる．NMT の分野では，He ら [2] が両方向の NMT モデルと各言語の言語モデルを用いて単言語データを相互に翻訳しあう双対学習を提案している．

Xia ら [3] は双対問題をモデルレベルで解く Model-Level Dual Learning を提案している．この手法は双対問題を解く単一のモデルを学習するもので，NMT であれば 1 つのモデルで双方向翻訳が可能になる．森下ら [4] や Chien ら [5] は，Xia らと同様のアーキテクチャを持つ Transformer ベースモデルの学習に Reconstruction や Dual Integration などの補助タスクを導入している．これらの補助タスクの導入により，モデルパラメータを増やすことなくモデル性能の改善を達成している．本稿では表記の簡潔さのため，双方向翻訳のために拡張された Transformer を Chien らに従い Dualformer と呼ぶ．

Dualformer の先行研究 [3] [4] [5] では，ドイツ語・英語，フランス語・英語，中国語・英語の 3 つの言語対が用いられており，いずれの言語対でも Transformer ベースモデルからの性能改善が見られる．しかし，日本語・英語の言語対での結果は示されていない．またモデルの評価はすべてドメイン内データで行っており，Dualformer のドメインシフトに対する頑健性は不明である．そのため本研究では，独英翻訳および日英翻訳の Dualformer をドメイン内・ドメイン外データで評価し，Dualformer のドメイン汎化性能を検証する．

2　Dualformer

本節では Dualformer の構成や翻訳処理，学習方法について述べる．

2.1　アーキテクチャ

Dualformer は，言語 X から言語 Y 方向翻訳の Transformer ベースモデルのデコーダ（図 1 右）と $Y \rightarrow X$ 方向翻訳 Transformer ベースモデルのデコーダ（図 1 左）から構成される．

$X \rightarrow Y$ 方向翻訳では，X 側デコーダは Cross Attention を無視して埋め込み層と Self Attention，全結合層だけ使うことで $X \rightarrow Y$ 方向のエンコーダとして振る舞わせる．この際，Self Attention への入力に対するマスク処理は無効にする．Y 側のデコーダ

はそのまま $X \to Y$ 方向のデコーダとして用いる．$Y \to X$ 方向翻訳も同様に，Cross Attention を無視した Y 側デコーダを $Y \to X$ 方向のエンコーダとして用い，X 側デコーダはそのまま $Y \to X$ 方向のデコーダとして使う．このように Dualformer はモデルの一部を Transformer のエンコーダもしくはデコーダとして使うことで双方向翻訳を実現する．

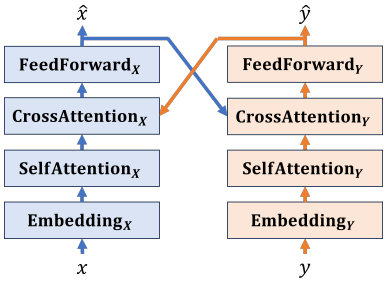

図 1　Dualformer 概略図

2.2　翻訳処理

まず，言語 X，Y のコーパスから得た対訳サンプルのトークン系列 x，y は埋め込み層により特徴ベクトル系列 $\mathbf{x}$，$\mathbf{y}$ に変換する（式 (1)，(2)）．

$$\mathbf{x} = \mathrm{Embed}_X(x) \tag{1}$$

$$\mathbf{y} = \mathrm{Embed}_Y(y) \tag{2}$$

Dualformer の翻訳処理では，モデルの一部を Transformer のエンコーダもしくはデコーダとして用いる．言語 X から言語 Y への翻訳における Dualformer のエンコーダ部を式 (3)，デコーダ部を式 (4) に示す．

$$E_{X\to Y}(\mathbf{x}) \triangleq \mathrm{FF}_X(\mathrm{SA}_X(\mathbf{x})) \tag{3}$$

$$D_{X\to Y}(\mathbf{h}^x, \mathbf{y}) \triangleq \mathrm{FF}_Y(\mathrm{CA}_Y(\mathbf{h}^x, \mathrm{SA}_Y(\mathbf{y}))) \tag{4}$$

ここで $\mathbf{h}^x$ は $X \to Y$ 方向翻訳のエンコーダの中間表現であり，式 (3) の処理の結果として得られる．また FF，CA，SA はそれぞれ全結合層，Cross Attention，Self Attention を表す．

式 (3)，(4) を用いると，Dualformer による $X \to Y$ 方向の翻訳処理はデコーダ出力分布 $\hat{\mathbf{y}}$ を用いて式 (5) のように書ける．

$$\hat{\mathbf{y}} = D_{X\to Y}(E_{X\to Y}(\mathbf{x}), \mathbf{y}) \tag{5}$$

同様に，$Y \to X$ 方向翻訳は式 (3)，(4) の X と Y を入れ替えると，$Y \to X$ 方向翻訳のデコーダ出力分布 $\hat{\mathbf{x}}$ を用いて式 (6) で表せる．

$$\hat{\mathbf{x}} = D_{Y\to X}(E_{Y\to X}(\mathbf{y}), \mathbf{x}) \tag{6}$$

$X \to Y$ および $Y \to X$ 方向の翻訳結果 $\hat{y}$，$\hat{x}$ は $\hat{\mathbf{y}}$，$\hat{\mathbf{x}}$ にそれぞれ Softmax 関数を適用して確率分布に直した後，各時刻で確率最大のトークンを抽出することで取得する．その処理を式 (7)，(8) にそれぞれ示す．

$$\hat{y} = \mathrm{argmax}(\mathrm{softmax}(\hat{\mathbf{y}})) \tag{7}$$

$$\hat{x} = \mathrm{argmax}(\mathrm{softmax}(\hat{\mathbf{x}})) \tag{8}$$

2.3　Reconstruction

本論文では，Dualformer の補助タスクとして Reconstruction を使用する．これは入力系列を再現するタスクである．なお，先行研究 [5] では Masked Language Model Loss と Dual Integration Loss も用いているが，本論文では使用しないため説明は省く．

Dualformer を用いた言語 X，Y の入力に対する Reconstruction は，Reconstruction のデコーダ出力分布 $\mathbf{x}'$，$\mathbf{y}'$ を用いて，それぞれ式 (9)，(10) で表される．

$$\mathbf{x}' = D_{Y\to X}(D_{X\to Y}(E_{X\to Y}(\mathbf{x}), \mathbf{y}), \mathbf{x}) \tag{9}$$

$$\mathbf{y}' = D_{X\to Y}(D_{Y\to X}(E_{Y\to X}(\mathbf{y}), \mathbf{x}), \mathbf{y}) \tag{10}$$

翻訳処理と同様に，Reconstruction 結果のトークン系列 x'，y' は式 (11)，(12) でそれぞれ求める．

$$x' = \mathrm{argmax}(\mathrm{softmax}(\mathbf{x}')) \tag{11}$$

$$y' = \mathrm{argmax}(\mathrm{softmax}(\mathbf{y}')) \tag{12}$$

2.4　目的関数

本論文では，Dualformer の目的関数として，これまで説明した翻訳と Reconstruction 処理に係る 2 つの目的関数の和を用いる．

翻訳の目的関数は式 (13) に示す．これは各方向の翻訳結果 $\hat{x}$，$\hat{y}$ の Cross Entropy 誤差の和である．同様に，式 (14) の Reconstruction の目的関数は，各方向の Reconstruction 結果 x'，y' の Cross Entropy 誤差の和である．

$$L_{\mathrm{TR}} = -\log p(\hat{y} \mid x) - \log p(\hat{x} \mid y) \tag{13}$$

$$L_{\mathrm{RC}} = -\log p(x' \mid x) - \log p(y' \mid y) \tag{14}$$

Dualformer の学習に用いられる目的関数は式 (13)，(14) を用いて式 (15) となる．

$$L = L_{\mathrm{TR}} + \alpha L_{\mathrm{RC}} \tag{15}$$

ここで，α は Reconstruction の使用割合である．この値が 0 のときは翻訳の目的関数のみで学習を行う．

3 実験設定

先行研究 [3] [4] [5] ではドメイン外データに対する評価を行っていない．また実験された言語対はドイツ語・英語，フランス語・英語，中国語・英語であり日本語・英語における性能は不明である．そのため，ドイツ語・英語および日本語・英語で Dualformer を構築し，ドメイン内とドメイン外データの両方に対する性能から Dualformer のドメインシフトに対する頑健性を検証する．

3.1 学習データ

独英翻訳および日英翻訳 Dualformer の訓練および評価データの構成を述べる．また，データの前処理やトークン化方法も示す．

3.1.1 データセット

独英翻訳モデルの学習に使う対訳データは，IWSLT'14 独英翻訳データセット（TED 公演ドメイン）の訓練セットである．評価には IWSLT'14 の開発，テストセットに加え，ドメイン外の WMT'17 独英翻訳タスク（ニュースドメイン）の開発，テストセットも使用する．

日英翻訳モデルの学習には，京都フリー翻訳タスク（Kyoto Free Translation Task, KFTT，京都関連 Wikipedia 記事ドメイン）の訓練セットを用いる．モデルの評価には KFTT に加えて，ドメイン外である ASPEC 日英翻訳データセット（学術論文ドメイン）の開発，テストも使用する．

独英翻訳および日英翻訳の学習・評価データの内訳は表 1 にそれぞれ示す．

表 1 実験で用いるデータセット（下線は訓練に使うデータセット）

言語対	データセット	サブセット [文数] 訓練	開発	テスト
独英	IWSLT'14	160,239	7,283	6,750
	WMT'17	-	45,901	3,003
日英	KFTT	440,288	1,166	1,160
	ASPEC	-	1,790	1,812

3.1.2 前処理

正規化

英語とドイツ語のデータには，小文字化を行った．日本語のデータには小文字化に加え，NFKC 正規化を行った．

単語分割

トークン化の前処理として単語分割を行う．ドイツ語と英語のデータは Moses [6] のトークナイザで単語に分割する．日本語のデータは MeCab を用いて分かち書きする．MeCab の辞書は NEologd [1] を用いる．

トークン化

本稿で使用する全言語でトークンは BPE（Byte Pair Encoding）[7] である．BPE の学習とテキストの BPE 化には Subword-NMT [2] を用いる．

独英翻訳と日英翻訳の BPE モデルはそれぞれドメイン内データの IWSLT'14 と KFTT の訓練セットに基づいて構築する．BPE の語彙サイズは，独英翻訳ではドイツ語が 8,844，英語が 6,628 であり，日英翻訳では日本語が 18,804，英語が 16,788 である．各言語対のドメイン外データである WMT'17 と ASPEC の BPE 化にもドメイン内データに基づく BPE モデルを用いる．

3.2 モデル設定

Dualformer およびベースラインの Transformer ベースモデルの構築と評価には fairseq [8] を用いる．

モデル構成は，アーキテクチャや言語対にかかわらず共通である．Dualformer のデコーダブロックはいずれも 6 層（Transformer の場合はエンコーダ・デコーダともに 6 層），Multi-Head Attention は 512 次元で 8 ヘッド，全結合層は 2,048 次元である．

3.3 評価方法

評価用のモデルは，訓練セットのドメインの開発セットへの損失（式 (15)）が最小のものを用いる．

評価指標は BLEU を用いる．BLEU の計算は，予測された BPE の系列を単語系列に直してから行う．単語列は，BPE 系列をデトークナイズした後，日本語の場合は MeCab，英語・ドイツ語の場合は Moses のトークナイザで単語分割して取得する．

4 実験結果

4.1 独英翻訳

表 2 に IWSLT'14 および WMT'17 に対するモデル性能を示す．ここで，各表の TF はベースラインの Transformer ベースモデルを表す．DF は Dualformer ベースモデルで，右下の数字は式 (15) における

1) https://github.com/neologd/mecab-ipadic-neologd
2) https://github.com/rsennrich/subword-nmt

表 2 IWSLT'14 で訓練した独英翻訳モデルの性能（下線はドメイン内データ，太字は最良の BLEU 値）

モデル	IWSLT14				WMT17			
	独 → 英		英 → 独		独 → 英		英 → 独	
	valid	test	valid	test	valid	test	valid	test
TF	32.22	31.14	-	-	12.55	14.91	-	-
	-	-	26.45	25.68	-	-	10.46	12.90
$DF_{0.0}$	32.46	31.26	26.81	25.68	13.06	15.44	10.87	13.42
$DF_{1.0}$	33.69	32.68	**27.65**	26.67	**14.30**	**17.07**	11.77	**14.27**
$DF_{1.25}$	**33.86**	**32.74**	**27.65**	**26.86**	14.25	16.82	**11.89**	14.11

表 3 KFTT で訓練した日英翻訳モデルの性能（下線はドメイン内データ，太字は最良の BLEU 値）

モデル	KFTT				ASPEC			
	日 → 英		英 → 日		日 → 英		英 → 日	
	valid	test	valid	test	valid	test	valid	test
TF	19.35	21.92	-	-	5.68	5.53	-	-
	-	-	15.41	16.14	-	-	4.43	4.09
$DF_{0.0}$	20.84	24.33	20.00	20.79	7.88	8.08	7.70	7.68
$DF_{0.5}$	**21.68**	**24.59**	**20.80**	**21.11**	**8.84**	**8.94**	**8.74**	**8.48**
$DF_{1.0}$	20.40	23.46	19.73	20.32	7.77	7.98	8.47	8.33

Reconstruction の使用割合 α である．

まず Transformer と Dualformer の結果を比較すると，Dualformer はドメイン内とドメイン外データのいずれに対しても Transformer を上回る性能を達成した．Dualformer の中でも，Reconstruction を使用したモデルはそうでないモデルよりも性能が高く，IWSLT'14 では $\alpha = 1.25$ のときが，WMT'17 では $\alpha = 1.0$ のときが最良の性能だった．

Dualformer の性能が Transformer を上回る理由は，Self Attention と全結合層がエンコーダの一部としてもデコーダの一部としても使われることで頑健なドメインマッピングが行われたためだと考えられる．また，一度に両方向の翻訳の学習を行うことで実質的な学習回数や使用データ数が 2 倍になり，Transformer よりも密度の高い学習が行われたことも理由として挙げられる．

次に Reconstruction の使用割合 α に注目すると，IWSLT'14 に対しては α を大きくすると翻訳性能が高くなり，WMT'17 に対しては逆に α を抑えたモデルの性能が高かった．これは，ドメイン内データを翻訳する際は Reconstruction によるドメイン知識の増強が有効だが，ドメイン外データの翻訳ではむしろドメイン知識が汎化性能を下げてしまうためだと考えられる．

4.2 日英翻訳

表 3 に KFTT に対するモデル性能および ASPEC に対するモデル性能を示す．

Dualformer は，全く異系統の言語対である日本語・英語の翻訳でもドメインの内外問わず Transformer の性能を超えた．特に英 → 日方向の性能向上は大きく，Transformer に対する $\alpha = 0.5$ の Dualformer の BLEU 増加率は，KFTT のテストセットでは日 → 英方向で 1.12，英 → 日方向で 1.31 であり，ASPEC のテストセットではそれぞれ 1.62 と 2.07 である．

この結果から，言語対にかかわらず Dualformer がドメインシフトへの頑健性を持つことが示された．

5 おわりに

本論文では単一モデルによる双方向翻訳が可能な Dualformer のドメイン汎化性能を調べた．実験では，Dualformer がドイツ語・英語および日本語・英語の両言語対において，ドメイン内とドメイン外データのいずれに対してもベースラインの Transformer の性能を上回った．この結果は，Dualformer は言語対にかかわらず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．

今後は，より大規模もしくは小規模なデータセットあるいは先行研究と本研究で使用していない言語対など異なる設定での Dualformer の性能を検証したい．また，翻訳以外の系列変換タスクに対しても Dualformer が有効なのか調査したい．

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 Plosukhin. Attention is all you need, 2017. https://arxiv.org/abs/1706.03762.

[2] Di He, Yingce Xia, , Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. In **30th Conference on Neural Information Processing System**, 2016.

[3] Yingce Xia, Xu Tan, Fei Tian, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Model-level dual learning. In **Proceedings of Machine Learning Research**, 2018.

[4] 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.

[5] Jen-Tzung Chien and Wei-Hsiang Chang. Dualformer: A unified bidirectional sequence-to-sequence learning. In **International Conference on Acoustics, Speech, and Signal Processing**, 2021.

[6] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In Sophia Ananiadou, editor, **Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions**, pp. 177–180, 2007.

[7] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, 2016.

[8] 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, 2019.