語彙切換に基づくニューラル機械翻訳の遠ドメイン適応

佐藤 翔悦 [*1]　佐久間 仁 [*1]　吉永 直樹 [*2]　豊田 正史 [*2]　喜連川 優 [*2*3]
[*1] 東京大学大学院 情報理工学系研究科
[*2] 東京大学 生産技術研究所　　[*3] 国立情報学研究所
{shoetsu, jsakuma, ynaga, toyoda, kitsure}@tkl.iis.u-tokyo.ac.jp

1 はじめに

ニューラル機械翻訳 (NMT) は，モデルの訓練時と異なるドメインにおいて大きくその性能が低下する [1, 2]. 現状，高精度のニューラル翻訳モデルを訓練できるほど十分な量の対訳コーパスが利用できるドメインは非常に少なく，多くのドメインでニューラル機械翻訳の高い精度を享受できていない．

この問題に対し，データが豊富に利用可能なドメイン（元ドメイン）から，データが少なく，実際に翻訳を行いたいドメイン（目標ドメイン）へのドメイン適応が盛んに研究されている．具体的には，元ドメインの大規模対訳コーパスで事前訓練されたモデルを目標ドメインの少数の対訳コーパスで再訓練する手法 (fine-tuning) [3, 4] と，2 つのドメインのデータを合わせて，モデルを同時に訓練する手法（マルチドメイン学習）[5, 6] が研究されている．しかしながら，これら既存の適応手法では，ドメイン適応で問題となる多様なドメイン間の差のうち，高頻度で出現するドメイン特有の文体や単語の翻訳は扱えるものの，低頻度の事象，特に語彙の違いについて，直接的に対処することができていない．実際のドメイン適応では十分な量の訓練データがあるドメインは少なく，出現する語彙や語義（の分布）が元ドメインと大きく異なる目標ドメイン（遠ドメイン）への適応が求められることを考慮すると，この点に対処することがドメイン適応の適応範囲を広げる上で重要となる．

ドメイン適応において，モデルの語彙がどのように設定され，それがどのような問題を引き起こすかを整理してみよう．まず fine-tuning では，元ドメインにおいて訓練されたモデルのパラメータを更新するため，目標ドメインでのみ出現する語は全て未知語となる．マルチドメイン学習では元ドメインと対象ドメイン両方を参照してモデルの語彙を設計できるため，この問題は軽減されるが，低頻度語については依然カバーできない．また，どちらの適応手法でも，ドメインによって語義（の分布）が変化する単語の扱いは難しい．例えば，“conductor” という単語が元ドメインで「指揮者」，目標ドメインで「導体」と訳されるとき，適切な訳を学習することは難しくなる．未知語についてはサブワードを利用して対処を測るのが標準的であるが，ドメインが異なるとサブワードが持つ語義（の分布）の違いのために，適切な翻訳が困難である．

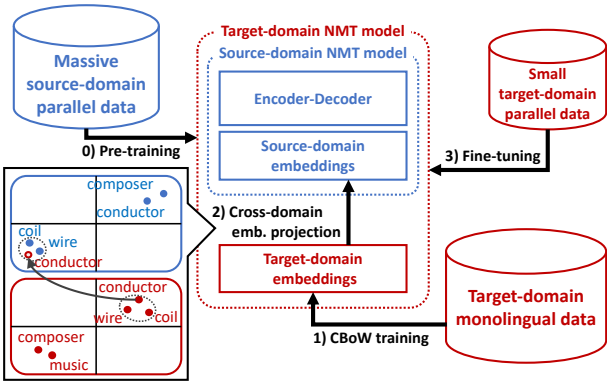

図 1: 語彙交換に基づく NMT モデルのドメイン適応．

そこで，本研究では目標ドメインにおける生コーパスは比較的容易に収集可能であると仮定した上で，訓練済みニューラル翻訳モデルの語彙とその単語埋め込みの切り換え（以降，**語彙切換**）によってドメイン間の語彙および語義（の分布）の差異を解決する．具体的には，目標ドメインの生コーパスから単語埋め込みを訓練し，元ドメインにおける訓練済みモデルの単語埋め込み空間へ写像を行う．その後，写像した単語埋め込みを訓練済みモデルの埋め込み層の初期値とした上で，目標ドメインにおける小規模な対訳コーパスを用いて fine-tuning を行う．本手法により，モデルの語彙およびその語義を目標ドメインに適したものに切り替えることが可能となる（図 1）．

実験では，元ドメインとして Japanese-English Subtitle Corpus (JESC) [7] を，目標ドメインとして Asian Scientific Paper Excerpt Corpus (ASPEC) [8] を用いた上で，英日翻訳タスクにおいて提案手法の評価を行う．実験の結果，標準的な fine-tuning に基づく手法と比較して最大で 18.4 ポイント，マルチドメイン学習と比較して最大 6.6 ポイントの BLEU の改善を確認した．

2 提案手法

本節では，NMT モデルにおけるドメイン間の語彙および語義の差異を解決する手法を提案する．提案手法は訓練済みモデルの埋め込み層にのみ作用するため，任意の深層学習モデルに利用可能であるが，本稿では

簡単のため，標準的な encoder-decoder モデルである Transformer [9] を仮定し，議論を進める．
提案手法の概要は以下の通りである．

Step 0 (Pre-training) 元ドメインの対訳コーパスから翻訳モデルを訓練する．モデルの語彙は元ドメインのコーパスから構築される．

Step 1 (Inducing target-domain embeddings) 原言語・目的言語について，目標ドメインの生コーパスから単語埋め込みを訓練する．本研究では翻訳タスクとの親和性を考慮し [10]，Continuous Bag-of-Words (CBoW) [11] を用いる．

Step 2 (Embedding projection) Step 1 で獲得した目標ドメインにおける原言語・目的言語の単語埋め込みを，Step 0 の訓練済みモデルの encoder，decoder の埋め込みの空間へとそれぞれ写像する．

Step 3 (Fine-tuning) Step 2 で写像した単語埋め込みをモデルの埋め込み層の初期値とし，目標ドメインの対訳コーパスを用いて fine-tuning を行う．初期化の際に，モデルの語彙集合もそれぞれの単語埋め込みと対応するものへと切り換える．

通常の fine-tuning との違いは，目標ドメインの生コーパスから訓練した単語埋め込みによって，fine-tuning の直前にモデルの埋め込み層および語彙を初期化する点のみである．そのため我々の提案手法は極めて低コストに適用可能であり，翻訳モデルの構成に依存しないことから，多くの既存手法と併用可能である．Step 2 における埋め込みの写像については，1) 多言語単語埋め込み獲得のための線形写像に基づく手法 [12] と，2) タスクに特化した単語埋め込み獲得のための非線形写像に基づく手法 [13] の 2 種類の写像手法を比較検討する．以下では，それぞれの手法について述べる．

2.1 線形写像に基づく語彙切換

1 つ目の手法として，多言語単語埋め込みの計算で用いられる，直交行列を用いた線形写像に基づく手法 [12] を用いる．本手法を含め，教師あり学習に基づく単語埋め込みの言語間写像では，単語単位の対訳辞書を用いて同じ意味を持つ単語の埋め込みが一致するように学習を行うことが多い．一方で，我々の目的は同言語内のドメイン間写像であり，多言語単語埋め込みにおける対訳辞書は利用できない．

そこで，本研究では線形写像の学習のための訓練データとして両ドメインに共通する単語を辞書として用いる．両ドメインに共通して存在する単語の中には語義が異なるものもあると考えられるが，種類数としては少なく写像の学習には悪影響を与えないと仮定する．また，両ドメインに共通する単語も目標ドメインで学習した埋め込みを写像してモデルの初期値とするため，目標ドメインにおける語義を適切に表現し，効果的に fine-tuning が行えると考える．

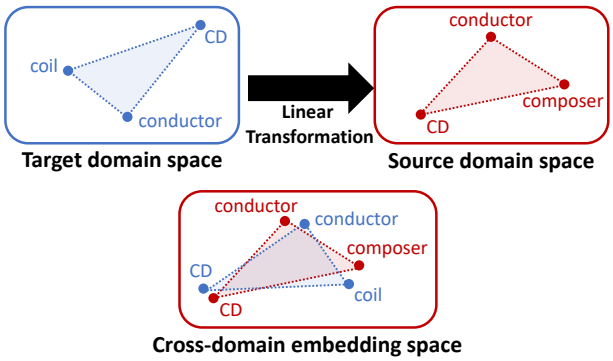

図 2: 埋め込みのドメイン間写像において，線形変換を用いた場合に考えられる問題の概要図．

2.2 非線形写像に基づく語彙切換

両言語の埋め込み空間の間でトポロジーの類似性が仮定できる言語間の単語埋め込みの写像と異なり，本研究で行うドメインおよびタスクを横断する写像では，埋め込み空間のトポロジーの違いが問題となる可能性がある．例えば，図 2 に示すように，ドメイン間で共通する単語の埋め込みを一致させるように直交行列を用いた線形変換を学習した場合，その写像は埋め込み空間全体に対する回転として実現される．そのため，“coil” という単語の埋め込みが，写像先であるモデルの埋め込み空間における “composer” の埋め込みに類似してしまうという問題が生じうる．

そこで 2 つ目の手法として，異なるタスク間で単語埋め込みの写像を学習する Locally Linear Mapping (LLM) [13] を用いた非線形的な埋め込みのドメイン間写像を提案する．LLM を用いた写像ではある単語と埋め込み空間上で近傍となるドメイン共通単語との局所的なトポロジーを保存するように，それぞれの単語ごとに写像を行うため，写像全体としては非線形的なものとなる．そのため，同じ表層を持つ単語であってもドメインごとの意味に応じた異なる埋め込みへの写像が容易になる．加えて，その写像全体としての非線形性から，上記の問題を解消することが可能であると考える．

3 実験

本研究では元ドメインとして JESC [7] を，目標ドメインとして ASPEC [8] を用いた英日翻訳タスクで評価を行う．翻訳モデルとしては，トークン化およびモデルの語彙の構築を単語単位で行うモデル (word-level) と，サブワード単位で行うモデル (subword-level) の 2 種類においてドメイン適応した結果を比較することで，サブワードの利用により未知語がほぼ生じない状況下でもなお提案手法が有効であるかの検証を行う．

En→Ja		JESC →	ASPEC
対訳文対数	training	2,797,388	2,000,000
	(fine-tuning)	-	100,000
	development	2,000	1,790
	testing	-	1,812
出現単語の種類数 (En)		161,695	637,377
出現単語の種類数 (Ja)		169,649	384,077
共通する単語の種類数 (En)		46,950 (7.4% in ASPEC)	
共通する単語の種類数 (Ja)		43,608 (11.4% in ASPEC)	

表 1: 各ドメインの対訳文対数及び出現単語種類数.

encoder/decoder 層	6	学習率 (初期)	1e-3
attention head の数	8	(warmup)	1e-7
Transformer の次元数	2048	バッチ内トークン数	4096
埋め込み次元数	512	訓練ステップ数 (training)	325k
語彙サイズ (word-level)	50k	(fine-tuning)	325k
(subword-level)	16k	Dropout 率	0.1

表 2: 翻訳モデルの主要なハイパーパラメータ.

3.1 実験設定

データセット 本実験のデータセットとして用いる JESC [7] および ASPEC [8] の性質とその前処理について述べる．JESC は映画・テレビ番組の字幕から構築されたコーパスである．一方で，ASPEC は科学技術に関する論文から構築されている．そのため，2 つのデータセットのドメインは大きく異なり，共有している語彙の割合も小さい（表 1）．どちらのデータセットもその分割は公式のものに準じた上で，ASPEC の training set については慣例に従い，先頭から 2,000,000 件の比較的高品質な翻訳対のみを用いた．fine-tuning set については，目標ドメインの training set から無作為に 100,000 件サンプルした．

前処理としては，Moses toolkit[1] (v4.0) および KyTea[2] (v0.4.2) をそれぞれ英語，日本語の単語分割に用いた後，英語については Moses toolkit を用いて truecasing を行った．この前処理済みデータセットに対し，SentencePiece[3] (v0.1.83) を適用してサブワードへと分割し，サブワード単位のモデルのためのデータセットとした．実験の再現性を考慮し，それぞれのドメインの training set 全てを擬似的な生コーパスとして用いた上で，各言語における単語埋め込み，サブワードの分割，サブワード埋め込みの訓練をドメインごとに行った．

モデル 翻訳モデルには fairseq (v0.8.0)[4] を用いて実装された Transformer [9] を採用し，Adam [14] で最適化を行った．主要なハイパーパラメータは表 2 の通りである．Transformer においては decoder の埋め込み層と出力層を共有することが一般的であり，本実験においてもその設定を採用した．その場合，提案手法の効果は出力層にも影響し，語彙切換によって元ドメインにおける未知語の出力も可能となる．

実験では，既存研究における提案手法を含む以下の 6 つの設定において訓練されたモデルを比較する．

[1] https://github.comd/moses-smt/mosesdecoder
[2] http://www.phontron.com/kytea
[3] https://github.com/google/sentencepiece
[4] https://github.com/pytorch/fairseq

Model	word	subword
No adaptation		
Out-domain	4.36	3.50
In-domain	10.09	11.06
Baselines		
Fine-tuning	5.30	11.01
Domain token mixing	17.98	17.53
Proposed		
Fine-tuning + Proposed (linear)	18.18	18.04
Fine-tuning + Proposed (LLM)	**23.65**	**24.14**

表 3: 目標ドメイン (ASPEC) における英日翻訳タスクの BLEU スコア．目的ドメインの翻訳対は 100,000 件.

Out-/In-domain 元ドメインの training set，または目標ドメインの fine-tuning set のみを用いてモデルを訓練する．

Fine-tuning 目標ドメインの fine-tuning set を用いて **Out-domain** モデルを再訓練する [3].

Domain token mixing 元ドメインの training set と目標ドメインの fine-tuning set を同時に用いてマルチドメイン学習を行う．その際に，出力文の先頭にドメインタグ（<src>または<tgt>）を加えたものを教師データとして与える [5]．この際，語彙の構築や埋め込みの訓練，SentencePiece の訓練には両ドメインの training set を合わせたものを擬似的な生コーパスとして用いる．

Fine-tuning + Proposed (linear/LLM) 2 節 参照．LLM の写像における近傍単語数は 10 に設定した．

これらの手法との比較によって，元ドメインの語彙を持ったモデルを fine-tuning する場合や，両ドメインのデータから語彙を設定しマルチドメイン学習を行う場合と比べ，目標ドメインに合わせた語彙切換を伴う fine-tuning が出力にどのような影響を与えるかを検証する．

3.2 実験結果

表 3 に単語単位のモデルおよびサブワードを用いた場合の結果を示す．まず単語単位のモデルについて，目標ドメインの小さな対訳コーパスのみを用いた場合 (**In-domain**) と比べて，**Out-domain** の結果は大きく劣り，異なるドメインにおけるモデルの性能の大きな低下が確認できた．また，ASPEC に加えて JESC の対訳コーパスも用いるモデルであるにも関わらず，**Fine-tuning** が **In-domain** を上回ることは無かった．この理由としては，語彙が大きく異なる遠ドメイン間では目標ドメインにおける低頻度語の多くが Out-of-Vocabulary となってしまい，fine-tuning の妨げとなったからであると考えられる．比較手法の中では，**Fine-tuning + Proposed (LLM)** が最も優れた結果を残し，同じく語

入力文	a low - density ablator layer was formed in the outside .
参照出力文	外側 に は 低 密度 の アブレータ 層 を 形成 し た 。
Out-domain	<unk> 密度 は 外側 に 形成 さ れ ま し た
In-domain	低 密度 層 で 覆 わ れ た 層 が 形成 さ れ た 。
Fine-tuning	<unk> の 外側 に 低 密度 の <unk> 層 が 形成 さ れ た 。
Domain token mixing	低 密度 で は ， 外部 の <unk> 層 が 形成 さ れ た 。
Fine-tuning + Proposed (linear)	低 い 電離 層 は 外側 に 形成 し た 。
Fine-tuning + Proposed (LLM)	内部 に は 低 密度 の アブレータ 層 が 形成 さ れ た 。

表 4: 表 3 左のそれぞれのモデルの ASPEC における出力例．下線部は元ドメイン (JESC) における未知語を表す．

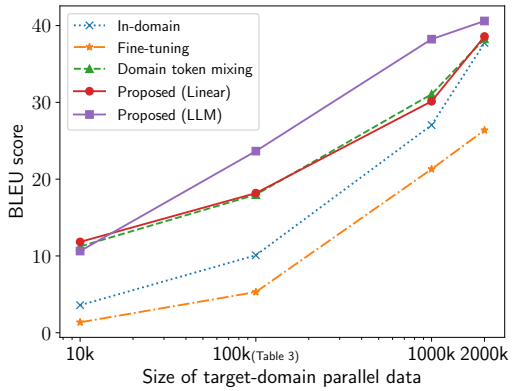

図 3: 目標ドメインの対訳コーパスのサイズと各モデルの BLEU スコア (word-level).

彙切換を行う提案手法である **Fine-tuning + Proposed (linear)** を大きく上回った．これは遠ドメイン適応における語彙切換の必要性と，2.2 節で述べた，埋め込み空間のトポロジーの違いを考慮した写像の有効性を示す結果であると言える．サブワードを用いた場合についても同様の傾向が確認され，未知語となるトークンがほぼ存在しないにも関わらず，**Fine-tuning + Proposed (LLM)** が他の手法を大きく上回った（表 3 右）．この理由としては，1) 適切なサブワードの分割はドメインによって異なること，2) 単語の語義がドメインごとに異なるように，同一の表層を持つサブワードの意味も異なり，その違いを提案手法により捉えることに成功したことが挙げられる．

また，単語単位のモデルにおいて，目標ドメインの訓練データサイズを変えた場合の結果を図 3 に示す．提案手法は一貫して **Fine-tuning** を上回る結果を示し，データサイズが大規模になった場合も同様であった．

表 4 に単語単位の翻訳モデルでの各手法（表 3 左）の出力例を示す．“ablator” と “アブレータ” は元ドメインにおける未知語である．比較手法のうち，**Fine-tuning + Proposed (LLM)** のみが “ablator” の翻訳に成功した．他の翻訳例においても同様の傾向が確認できた．

4 おわりに

本研究では，機械翻訳タスクにおけるドメイン適応の際，ドメイン間の語彙の差異が引き起こす問題に注目し，目標ドメインの生コーパスから訓練した単語埋め込みをモデルの埋め込み空間に写像して fine-tuning の際に用いることでその解決を試みた．また，線形写像による埋め込みのドメイン間写像における問題点を考察し，LLM による非線形写像がより適している事を示した．英日翻訳における実験の結果，LLM を用いた提案手法による大きな性能の改善を確認した．今後の課題として，本提案手法はその性質上広範なモデルやドメイン適応手法と併用可能であることから，機械翻訳のみならず対話応答や文書要約など，様々なテキスト生成タスクへの応用を検討している．

謝辞 本研究の一部は JSPS 科研費 19J14522，2019 年度国立情報学研究所 CRIS 委託研究，および JST，CREST，JPMJCR19A4 の支援を受けたものである。

参考文献

[1] P. Koehn and R. Knowles. Six challenges for neural machine translation. In *Proc. WNMT*, pp. 28–39, 2017.

[2] C. Chu and R. Wang. A survey of domain adaptation for neural machine translation. In *Proc. of COLING*, pp. 1304–1319, 2018.

[3] M.-T. Luong and C. D. Manning. Stanford neural machine translation systems for spoken language domains. In *Proc. IWSLT*, pp. 76–79, 2015.

[4] A. Bapna and O. Firat. Simple, scalable adaptation for neural machine translation. In *Proc. EMNLP-IJCNLP*, pp. 1538–1548, 2019.

[5] D. Britz, Q. Le, and R. Pryzant. Effective domain mixing for neural machine translation. In *Proc. WMT*, pp. 118–126, 2017.

[6] R. Wang, M. Utiyama, L. Liu, K. Chen, and E. Sumita. Instance weighting for neural machine translation domain adaptation. In *Proc. EMNLP*, pp. 1482–1488, 2017.

[7] R. Pryzant, Y. Chung, D. Jurafsky, and D. Britz. JESC: Japanese-English subtitle corpus. In *Proc. LREC*, 2018.

[8] T. Nakazawa, M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi, and H. Isahara. ASPEC: Asian scientific paper excerpt corpus. In *Proc. LREC*, pp. 2204–2208, 2016.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in NIPS*, pp. 5998–6008, 2017.

[10] M. Neishi, J. Sakuma, S. Tohda, S. Ishiwatari, N. Yoshinaga, and M. Toyoda. A bag of useful tricks for practical neural machine translation: Embedding layer initialization and large batch size. In *Proc. of WAT*, pp. 99–109, 2017.

[11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in NIPS*, pp. 3111–3119, 2013.

[12] C. Xing, D. Wang, C. Liu, and Y. Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proc. NAACL*, pp. 1006–1011, 2015.

[13] J. Sakuma and N. Yoshinaga. Multilingual model using cross-task embedding projection. In *Proc. CoNLL*, pp. 22–32, 2019.

[14] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proc. ICLR*, 2015.