言語横断類似度推定のための多言語文符号化器のドメイン適応

山内 洋輝 [1]　梶原 智之 [1]　桂井 麻里衣 [2]　二宮 崇 [1]
[1] 愛媛大学　[2] 同志社大学
{yamauchi@ai., kajiwara@, ninomiya@}cs.ehime-u.ac.jp　katsurai@mm.doshisha.ac.jp

概要

本研究では，多言語文符号化器のドメイン適応に取り組む．高度な専門知識が要求される医療や学術のドメインにおいては，目的ドメインに特化した事前訓練の有効性が知られている．しかし，様々な言語においてドメイン特化の単言語事前訓練モデルの開発が進む一方，言語横断情報検索などに応用可能な多言語モデルは存在しない．また，各言語における目的ドメインのコーパスを整備し多言語の事前訓練を行うには大きなコストが必要となる．そこで我々は，既存の多言語文符号化器および 2 言語の各々における目的ドメインに特化した単言語文符号化器を用いて，効率的にドメイン特化の言語横断文符号化器を開発する．3 つのドメインおよび言語対における翻訳ランキングの評価実験の結果，ドメイン適応なしのベースラインや既存のドメイン適応手法と比べて，提案手法の有効性を確認できた．

1　はじめに

Web 上に蓄積された膨大なデータから網羅的に情報を得るために，言語横断情報検索 [1] の技術が期待されている．埋め込みベースの言語横断情報検索などの応用を目指して，多言語文符号化器 [2–4] の研究も盛んに進められている．これらの多言語文符号化器は Wikipedia や CC100 [5] などの一般的なテキストを用いて訓練されているが，高度な専門知識が要求される医療や学術のドメインに対応するためには，目的ドメインに特化した多言語文符号化器を開発することが望ましい．しかし，これらのドメインに特化した単言語文符号化器 [6–11] の開発が進む一方で，言語横断情報検索に応用可能なドメイン特化の多言語文符号化器は存在しない．

そこで本研究では，ドメイン特化の言語横断文符号化器を構築する手法を提案する．ただし，各言語における目的ドメインのコーパスを整備して多言語の事前訓練を行うには大きなコストが必要となるため，汎用的な多言語文符号化器 [3,4,12,13] の開発と同様に，事前訓練済みの文符号化器を対訳コーパス上でファインチューニングするアプローチを採用する．先行研究では，知識蒸留 [3]，翻訳ランキング [4]，対照学習 [12]，敵対的学習 [13] などの手法を用いて，対訳文における原言語と目的言語の埋め込みを近づけることに焦点を当ててきた．一方で本研究では，原言語と目的言語の埋め込みを近づけるだけでなく，ドメイン知識の獲得も同時に行う．

学術（英日）・医療（英仏）・金融（英中）の 3 つのドメインおよび言語対における評価実験の結果，提案手法がドメイン適応なしのベースラインや既存のドメイン適応手法と比べて高い性能を達成することを確認できた．詳細な分析の結果，提案手法は目的ドメインの対訳コーパスが数千文対しか使用できない状況でも有効であることが明らかになった．

2　提案手法

図 1 に示すように，本研究では目的ドメインの対訳コーパス上でのファインチューニングを通して，多言語文符号化器から言語非依存の埋め込み（意味表現）を抽出しつつ，各言語におけるドメイン特化の単言語文符号化器からドメイン知識を蒸留する．このように獲得した意味表現（図 1 の黄色）は，目的ドメインに適した意味空間上で，入力言語に関わらず，意味的に類似した文に類似したベクトルを与えられる．本手法は，多言語文符号化器から得られる埋め込みを言語固有の情報を持つ言語表現と言語非依存の情報を持つ意味表現に分離する DREAM [12] および MEAT [13] から着想を得て言語非依存の埋め込みを訓練するとともに，Multilingual SentenceBERT [3] の知識蒸留から着想を得てドメイン特化の埋め込みを訓練するものである．

先行研究 [12,13] と同様に，多言語文符号化器から得られる埋め込みを，MLP_{*L} および MLP_{*M} の 2 つの多層パーセプトロンによって言語表現および意味表現に分離する．これを原言語 S および目的言語

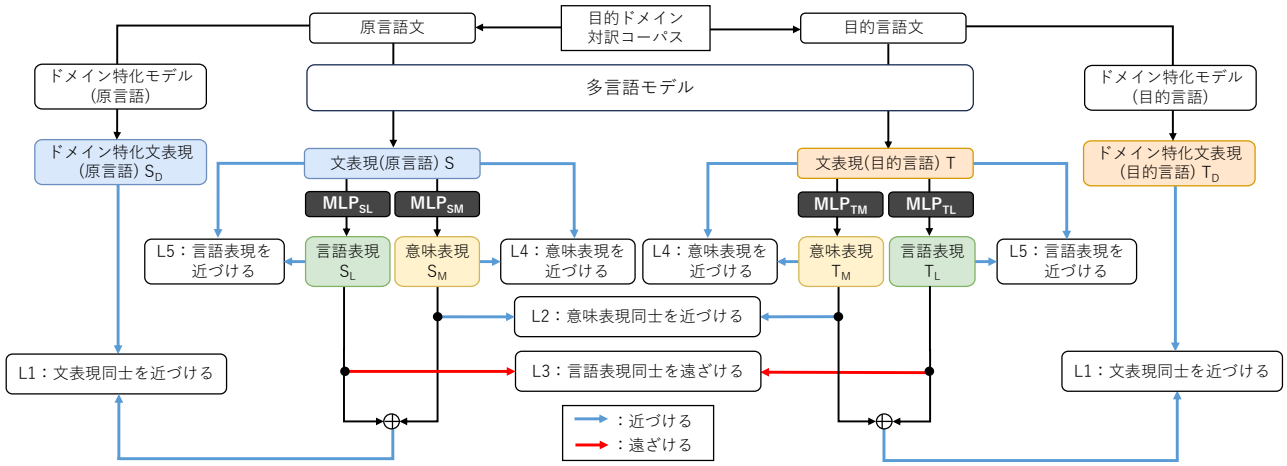

図 1　提案手法の概要図．目的ドメインの対訳コーパス，多言語文符号化器，各言語におけるドメイン特化の単言語文符号化器を用いて，多言語文符号化器から言語非依存なドメイン特化の意味表現を抽出するように MLP を訓練する．

T の両方で実施し，合計 4 つの MLP を以下の 5 つの損失関数に基づくマルチタスク学習で訓練する．

$$L = L_1 + L_2 + L_3 + L_4 + L_5 \tag{1}$$

2.1　ドメイン知識の蒸留

言語表現と意味表現を足し合わせた文表現を，ドメイン特化の単言語文符号化器から得た文表現に近づけることで，意味表現へのドメイン知識の蒸留を狙う．以下のように損失を定義する．

$$L_1 = 2 - (\cos((S_L + S_M), S_D) + \cos((T_L + T_M), T_D)) \tag{2}$$

ここで，$S_L \in \mathbb{R}^d$ および $T_L \in \mathbb{R}^d$ は原言語および目的言語の言語表現，$S_M \in \mathbb{R}^d$ および $T_M \in \mathbb{R}^d$ は各言語の意味表現，$S_D \in \mathbb{R}^d$ および $T_D \in \mathbb{R}^d$ は各言語のドメイン特化モデルから得た文表現を表す．ただし，d は埋め込みの次元数である．

2.2　言語間で意味表現同士を近づける

言語に依存せず，意味的に等価な文同士の意味表現を近づけたい．これを達成するために，対訳文の間で意味表現を近づける以下の損失を定義する．

$$L_2 = 1 - \cos(S_M, T_M) \tag{3}$$

2.3　言語間で言語表現同士を遠ざける

意味に依存せず，異なる言語の文同士の言語表現を遠ざけたい．これを達成するために，対訳文の間で言語表現を遠ざける以下の損失を定義する．

$$L_3 = \max(0, \cos(S_L, T_L)) \tag{4}$$

2.4　元の文表現と離れすぎない

言語非依存の意味表現を獲得するために，言語固有であるドメイン特化の単言語文符号化器に影響を受けすぎないようにしたい．これを達成するために，意味表現と多言語文符号化器から得られる元の文表現の大きな乖離を防ぐ以下の損失を定義する．

$$L_4 = 2 - (\cos(S, S_M) + \cos(T, T_M)) \tag{5}$$

意味表現と同様に，言語表現も元の文表現との大きな乖離を避けるために，以下の損失を定義する．

$$L_5 = 2 - (\cos(S, S_L) + \cos(T, T_L)) \tag{6}$$

2.5　実装の詳細

文表現には，各モデルから出力されるトークン埋め込みの平均プーリングを用いる．MLP には 1 層の順伝播型ニューラルネットワークを用いる．また，対訳コーパス上でファインチューニングするのは MLP のみであり，文符号化器は更新しない．

3　評価実験

3 つのドメインおよび言語対を対象とする翻訳ランキングタスクにおいて提案手法の性能を評価する．翻訳ランキングは，ある原言語文に対して意味的類似度の降順に目的言語文をランキングするタスクであり，対訳文を 1 位にランク付けできた割合を評価する ExactMatch および上位 10 件のうち何位に対訳文をランク付けできたかを評価する平均逆順位 MRR@10 によって自動評価する．本実験では，余弦類似度によって意味的類似度を推定する．

表 1 実験結果

		学術 (En-Ja)		医療 (En-Fr)		金融 (En-Zh)	
		ExactMatch	MRR@10	ExactMatch	MRR@10	ExactMatch	MRR@10
En → XX	LaBSE	0.908	0.927	0.943	0.949	0.476	0.537
(1)	L_6	0.000	0.001	0.000	0.002	0.000	0.002
(2)	$L_6 + L_7$	0.619	0.699	0.847	0.876	0.223	0.303
(3)	$L - L_1 + L_8$	0.916	0.934	**0.951**	**0.954**	0.503	0.565
	提案手法	**0.919**	**0.937**	**0.951**	**0.954**	**0.510**	**0.574**
XX → En	LaBSE	0.889	0.910	0.937	0.944	0.353	0.416
(1)	L_6	0.000	0.001	0.000	0.002	0.000	0.002
(2)	$L_6 + L_7$	0.493	0.584	0.898	0.915	0.053	0.002
(3)	$L - L_1 + L_8$	0.908	0.926	**0.949**	**0.953**	0.452	0.517
	提案手法	**0.911**	**0.929**	**0.949**	**0.953**	**0.462**	**0.528**

表 2 対訳コーパスの文対数

	訓練	検証	評価
学術（英日）	100,000	5,000	5,000
医療（英仏）	100,000	5,000	5,000
金融（英中）	50,000	5,000	5,000

3.1 実験設定

データセット 学術・医療・金融の 3 つのドメインにおいて実験した．学術ドメインにおいては，科研費の採択課題名[1]に関する英日の対訳コーパスを用いた．医療ドメインにおいては，WMT16 [14] の Biomedical Translation タスクにて採用された PubMed に関する英仏の対訳コーパスを用いた．金融ドメインにおいては，Financial Times の記事タイトルを抽出した英中の対訳コーパス [15] を用いた．これらの対訳コーパスの文対数を表 2 に示す．それぞれ，検証用および評価用に 5,000 文対ずつを無作為抽出し，残りを訓練用に用いた．

モデル ドメイン特化の単言語文符号化器には，学術ドメインにおいては英語の SciBERT[2] [6] および日本語の AcademicRoBERTa[3] [10]，医療ドメインにおいては英語の Bio_ClinicalBERT[4] [7] およびフランス語の DrBERT[5] [11]，金融ドメインにおいては英語の FinBERT[6] [8] および中国語の Mengzi-BERT[7] [9] を用いた．多言語文符号化器には，機械翻訳の品質推定のために言語表現と意味表現を分離する先行研究 [12,13] と同様に，最先端の多言語文符号化器のひとつである LaBSE[8] [4] を用いた．バッチサイズを 512，最適化手法を Adam [16]，学習率を 1e-5 として HuggingFace Transformers [17] を用いて訓練した．検証用データにおける式 (1) の損失が 3epoch 改善しない場合に訓練を終了した．

比較手法 言語表現と意味表現を分離しつつドメイン知識の蒸留を行う提案手法の有効性を評価するために，LaBSE の文表現をそのまま用いるベースラインに加えて，3 つの比較手法の実験を行う．まず，ドメイン知識の蒸留のみを行う比較手法 (1) として，言語表現と意味表現に分離する MLP_L および MLP_M の代わりに，原言語用の MLP_S および目的言語用の MLP_T を用いて，以下の損失 L_6 を考える．

$$L_6 = 2 - (\cos(S_S, S_D) + \cos(T_T, T_D)) \quad (7)$$

次に，言語表現と意味表現を分離しない比較手法 (2) として，ドメイン知識の蒸留に加えて原言語の表現と目的言語の表現を近づける損失 $L_6 + L_7$ を考える．

$$L_7 = 1 - \cos(S_S, T_T) \quad (8)$$

最後に，ドメイン知識の蒸留をしない比較手法 (3) として，言語表現と意味表現を足し合わせて元の文表現を復元する損失 $L - L_1 + L_8$ を考える．

$$L_8 = 2 - (\cos((S_L + S_M), S) + \cos((T_L + T_M), T)) \quad (9)$$

1) https://kaken.nii.ac.jp/
2) https://huggingface.co/allenai/scibert_scivocab_uncased
3) https://huggingface.co/EhimeNLP/AcademicRoBERTa
4) https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT
5) https://huggingface.co/Dr-BERT/DrBERT-7GB
6) https://huggingface.co/ProsusAI/finbert
7) https://huggingface.co/Langboat/mengzi-bert-base-fin
8) https://huggingface.co/sentence-transformers/LaBSE

表 3　アブレーション分析

					ExactMatch	
	L2	L3	L4	L5	En → Ja	Ja → En
LaBSE					0.908	0.889
(a)			✓	✓	0.867	0.820
(b)	✓	✓			0.010	0.419
(c)		✓	✓	✓	0.908	0.888
(d)	✓		✓	✓	0.919	**0.911**
(e)	✓	✓		✓	**0.920**	0.893
(f)	✓	✓	✓		0.918	**0.911**
提案手法	✓	✓	✓	✓	0.919	**0.911**

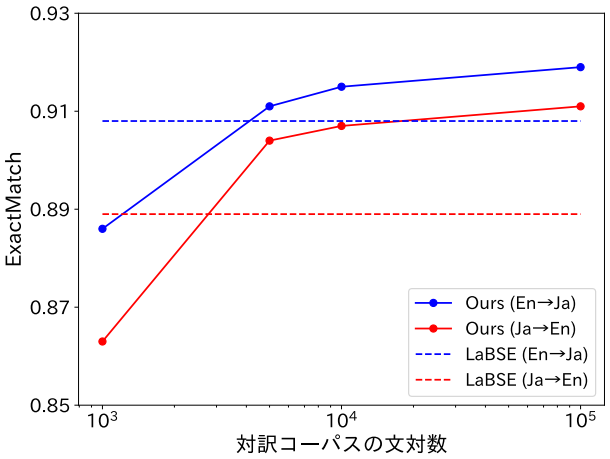

図 2　対訳コーパスの規模と言語横断情報検索の品質

3.2　実験結果

表 1 に実験結果を示す．上段は英語の文をクエリとして他言語の文を検索する場合の実験結果であり、下段は他言語の文をクエリとして英語の文を検索する場合の実験結果である．

LaBSE と比較して，全てのドメインおよび言語対において，提案手法が一貫して高い性能を達成できた．ドメイン知識の蒸留のみを行う比較手法 (1) は，対訳文を検索する能力が完全に失われており，単純なドメイン適応では言語横断モデルを訓練できないことがわかる．また，言語表現と意味表現を分離しない比較手法 (2) も著しく性能が劣化することから，言語非依存な意味表現を抽出することが言語横断の類似度推定のために重要であることがわかる．ドメイン知識の蒸留をしない比較手法 (3) は，元の LaBSE の性能を上回ったが，全てのドメインおよび言語対において提案手法以下の性能に留まった．そのため，言語非依存な意味表現を抽出するだけでも言語横断情報検索の性能改善を期待できるものの，ドメイン適応によってその品質をさらに改善できると言える．

3.3　アブレーション分析

提案手法のうち，ドメイン適応を司る L_1 以外の損失についての様々な組み合わせを評価した結果を表 3 に示す．紙面の都合上，学術ドメインにおける英日データの結果のみを掲載するが，他のドメインや言語対においても同様の結果が見られた．

(a) に性能悪化が見られるため，高い言語横断性能を達成するためには言語表現と意味表現を分離する L_2 および L_3 の損失が重要であることがわかる．(b) より，元の文表現を大きく更新することを防ぐ L_4 および L_5 を除くことで著しい性能悪化が見られることから，安定したドメイン適応の訓練のためにこれらの損失が重要であることがわかる．(c) および (e) から，意味表現に関する損失を除外することで，提案手法と比べて to-English 方向の性能が悪化することがわかる．一方で，(d) および (f) から，言語表現に関する損失を除外しても大きな影響はない．

3.4　対訳コーパスの規模に関する分析

訓練に使用する目的ドメインの対訳コーパスの規模を減らしつつ性能の変化を分析した結果を図 2 に示す．5,000 文対の対訳コーパスを用いた訓練でさえ，提案手法（実線）は LaBSE の性能（点線）を上回ることが明らかになった．特定のドメインにおける大規模な対訳コーパスを整備するには大きなコストがかかるため，小規模なデータで言語横断のドメイン適応を実現できることには価値がある．

4　おわりに

本研究では，多言語文符号化器のドメイン適応に取り組んだ．提案手法は，目的ドメインの対訳コーパスと目的ドメインに特化した単言語文符号化器を用いて，多言語文符号化器から言語非依存の意味表現を抽出しつつ，ドメイン知識を蒸留する．学術・医療・金融の 3 ドメインおよび英日・英仏・英中の 3 言語対における実験結果は，ドメインや言語に依存せず一貫して，提案手法が言語横断類似度推定の性能を改善できることを示した．本手法は，目的ドメインにおける数千文対の対訳コーパスからでさえ，多言語文符号化器のドメイン適応を実現できる．

謝辞

本研究は JSPS 科研費（基盤研究 B，課題番号：JP20H04484）および国立研究開発法人情報通信研究機構（NICT）の委託研究（課題番号：22501）による助成を受けたものです.

参考文献

[1] Mikel Artetxe and Holger Schwenk. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. **Transactions of the Association for Computational Linguistics**, Vol. 7, pp. 597–610, 2019.

[2] 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, 2020.

[3] Nils Reimers and Iryna Gurevych. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**, pp. 4512–4525, 2020.

[4] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT Sentence Embedding. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics**, pp. 878–891, 2022.

[5] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 4003–4012, 2020.

[6] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A Pretrained Language Model for Scientific Text. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing**, pp. 3615–3620, 2019.

[7] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly Available Clinical BERT Embeddings. In **Proceedings of the 2nd Clinical Natural Language Processing Workshop**, pp. 72–78, 2019.

[8] Dogu Araci. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. **arXiv:1908.10063**, 2019.

[9] Zhuosheng Zhang, Hanqing Zhang, Keming Chen, Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming Zhou. Mengzi: Towards Lightweight yet Ingenious Pre-trained Models for Chinese. **arXiv:2110.06696**, 2021.

[10] Hiroki Yamauchi, Tomoyuki Kajiwara, Marie Katsurai, Ikki Ohmukai, and Takashi Ninomiya. A Japanese Masked Language Model for Academic Domain. In **Proceedings of the Third Workshop on Scholarly Document Processing**, pp. 152–157, 2022.

[11] Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics**, pp. 16207–16221, 2023.

[12] Nattapong Tiyajamorn, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. Language-agnostic Representation from Multilingual Sentence Encoders for Cross-lingual Similarity Estimation. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 7764–7774, 2021.

[13] Yuto Kuroda, Tomoyuki Kajiwara, Yuki Arase, and Takashi Ninomiya. Adversarial Training on Disentangling Meaning and Language Representations for Unsupervised Quality Estimation. In **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 5240–5245, 2022.

[14] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 Conference on Machine Translation. In **Proceedings of the First Conference on Machine Translation**, pp. 131–198, 2016.

[15] Nicolas Turenne, Nicolas Turenne, Ziwei Chen, Guitao Fan, Jianlong Li, Yiwen Li, Siyuan Wang, and Jiaqi Zhou. Mining an English-Chinese Parallel Dataset of Financial News. **Journal of Open Humanities Data**, Vol. 8, No. 9, pp. 1–12, 2022.

[16] Diederik P. Kingma and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. In **Proceedings of the 3rd International Conference for Learning Representations**, 2015.

[17] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45, 2020.