Multilingual BERTの二言語領域適応に基づく対訳文同定*

飯田 頌平† 三好 健悟† 崔 鴻翌† 洪 博軒† 宇津呂 武仁† 永田 昌明‡
†筑波大学大学院 システム情報工学研究科 ‡NTT コミュニケーション科学基礎研究所

1 はじめに

ニューラル機械翻訳の翻訳精度は対訳文の質と量に大きな影響を受けるため，高品質な対訳文を自動的に獲得する研究が行われている [8, 17, 18]．また近年では Multilingual-BERT(M-BERT) [3] や LASER [1] をはじめとした，汎用的な事前訓練済みの多言語エンコーダが公開されており，これらの文埋め込みを用いて対訳文を同定する手法が提案されている [2]．しかし文埋め込みによる対訳文同定手法では，対訳文を獲得したい領域に適応したモデルをあらかじめ用意する必要があり，たとえば文献 [2] ではネパール語から英語の対訳文を得るためにネパール語・英語対訳コーパスを用いて領域適応した LASER を用いている．このように，質の高い対訳文を得るためには，対訳文同定モデルを領域適応するための対訳文が必要になるという問題がある．

本論文では，日英特許文書の領域において，二言語の非対訳特許文書から Masked Language Model (MLM) を訓練することによって，Multilingual-BERT の文埋め込みを日英特許文書に対して二言語領域適応する．そして，二言語領域適応された Multilingual-BERT の文埋め込みを，さらに二値分類器として fine-tuning することにより，日英特許文書間の対訳文同定を行う手法を提案する．提案手法においては，MLM による領域適応においては対訳コーパスを用いる必要がない点と，文対応手法 [11] を利用した正例，LASER を利用した負例という高品質な fine-tuning 用データを抽出して用いた点で前述の問題に対処した．モデルの評価において，二言語領域適応の有無の比較を行い，二言語領域適応の有用性を確認した．

*Identifying Parallel Sentences by Bilingual Domain Adaptation of Multilingual BERT

†Shohei Iida, Kengo Miyoshi, Hongyi Cui, Po-Hsuan Hung, Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba
‡Masaaki Nagata, NTT Communication Science Laboratories, NTT Corporation, Japan

2 関連研究

高品質な対訳文を得ることは機械翻訳において重要な問題であるため，長年にわたり様々なアプローチから取り組まれてきた．

ニューラル言語モデルが普及する前には，文献 [11, 12] において対訳辞書と統計量を用いて文対応を取る手法が提案された．

近年では文埋め込みベクトルを用いた手法が主流となっており，文献 [1] においては LASER [1] の文埋め込みにより Building and Using Comparable Corpora (BUCC) タスク [17, 18] における state-of-the-art を達成した．さらに文献 [2] では LASER [1] の文埋め込みをはじめ，複数の手法[1] を組み合わせることで WMT19 のパラレルコーパスフィルタリングタスク[2] における state-of-the-art を達成した．そのほか，文献 [4] においては，独立した二種類の Transformer [13] エンコーダにより原言語文の文埋め込みと目的言語文の文埋め込みを取得し，両者の内積による文類似度を計算して対訳文を同定する Dual-encoder アーキテクチャが提案された．さらに，文献 [16] においては，二値分類器として fine-tuning した M-BERT [3] によって Dual-encoder の出力を再計算する手法が提案された．

また，言語横断文書分類タスクにおいては，単言語で領域適応した二言語分の単語埋め込みを学習し，その後言語間の変換行列を学習することで得た二言語単語埋め込みによる領域適応を行う手法 [5] が提案されたが，本論文では M-BERT において二言語の Masked Language Model を再訓練することにより，単語単位ではなく文単位の二言語領域適応を実現している．

3 日英特許文書データ

本論文では，日本，米国の 2004 年の特許データからパテントファミリーに基づいて対訳文書対を抽出した特許文書データを用いる．日英特許文書データにおいては，文対応手法 (以下，内山法 [11] と呼ぶ) により

[1] LASER のほか，Zipporah [15]，Bicleaner [10]，Dual Conditional Cross-Entropy [7] が用いられた．

[2] http://www.statmt.org/wmt19/parallel-corpus-filtering.html

表 1: fine-tuning に使用する正例および負例の例

手法	文例
日本語文	歯つきのシステムにより、患者は、各ロッククリップ 82 について所望の角度を容易に設定し、再現することが可能である。
正例	The toothed system allows the patient to easily set and reproduce the desired angle for each locking clip 82.
負例 (LASER の出力)	For example, the patient may adjust both locking clips 82 to have the same angle with respect to the yokes 92.
負例 (無作為)	Examples of the plastic include nylon, acetal, polycarbonate, and polypropylene.

日本語文と英語文の対応が取られたセグメント対の列に区切られる．内山法による対訳文同定法では，文書の冒頭では比較的高精度な対応が取れているものの，末尾に近づくにつれて精度が低くなるという特徴がある．各セグメント対には文対応スコアが与えられるが，文対応スコアが高いものは文対応の精度がよく，逆に文対応スコアが低いものは文対応の精度が悪い傾向にある．文対応の精度がよいものは表 2 に示すように fine-tuning に，精度の悪いものは表 3 に示すように検証に用いた．

表 2: 二言語領域適応・対訳文分類器の訓練用日英特許文書データ

(a) 使用する日英特許文書データ

	文書数	セグメント対数	文数
日本語	240	97,417	113,716
英語	240	97,417	111,531

(b) 各手順で用いた訓練文

		文数・文対数
MLM 訓練文 (日+英)		225,247 文
fine-tuning	正例	約 18 万文対
	負例 (LASER)	約 90 万文対
	負例 (無作為)	約 90 万文対
	合計	1,986,492 文対

表 3: 対訳文分類器の開発・検証用日英特許文書データ

(a) 使用する日英特許文書データ

	文書数	セグメント対数	文数
日本語	1	799	963
英語	1	799	852

(b) スコアが閾値以下の 45 セグメント対から選定した開発文対・検証文対

	文数・文対数
開発文対	100 文対
検証文対	137×80 文対

4 Masked Language Model の再訓練による二言語領域適応

BERT [3] では，入力文のトークンを一定確率で [MASK] トークンで置き換え，置き換えられたトークンを予測する問題により Masked Language Model(MLM) が訓練されている．訓練済みの BERT はオープンドメインなデータによる MLM であるため，汎用性に富み様々なタスクに転用できる一方で特定の領域における表現能力には改善の余地がある．

そこで本論文では，日英二言語の特許データ用いて MLM の再訓練を行うことで，日英特許文書の対訳文同定タスクに特化したモデルを得る．

実験において，HuggingFace [3] の実装を使用し，240 特許文書から得た日本語特許文 113,716 文と英語特許文 111,531 文による計 225,247 文の特許文 (表 2(b)) を BertTokenizer によるトークン化処理の後に 2 エポック学習した [4]．また訓練の初期値に用いた M-BERT では，119,547 語彙の辞書を用いた 104 言語で訓練が実施され，12 ヘッドのマルチヘッド注意を持つ 12 層のエンコーダを持ち，各隠れ層は 768 次元であり，約 1.1 億のパラメータを有する．

5 LASER 文埋め込みを用いた言語横断文間距離

LASER [1] は英語と 93 言語の言語対からなる対訳コーパスから訓練された双方向 LSTM モデルのエンコーダであり，言語にとらわれない文埋め込みを取得できる．本論文では検証文集合内の 137 文の日本語文と 80 文の英語文 (表 3(b)) のすべての組み合わせで文間距離を比較し，もっとも文間距離が近い日本語文と英語文の文対を対訳文として同定する．

本論文では，fine-tuning の負例作成，および提案法との比較のための性能評価において LASER を用いた．実験においては，日本語文では MeCab [5]，英語文では Moses Tokenizer [9] によるトークン化処理の後，

[3] https://github.com/huggingface/transformers

[4] NVIDIA Tesla P100 を二枚使用し，バッチサイズを 8 に設定したところ，1 エポックあたりの訓練時間は約 30 分であった．

[5] http://mecab.sourceforge.net/

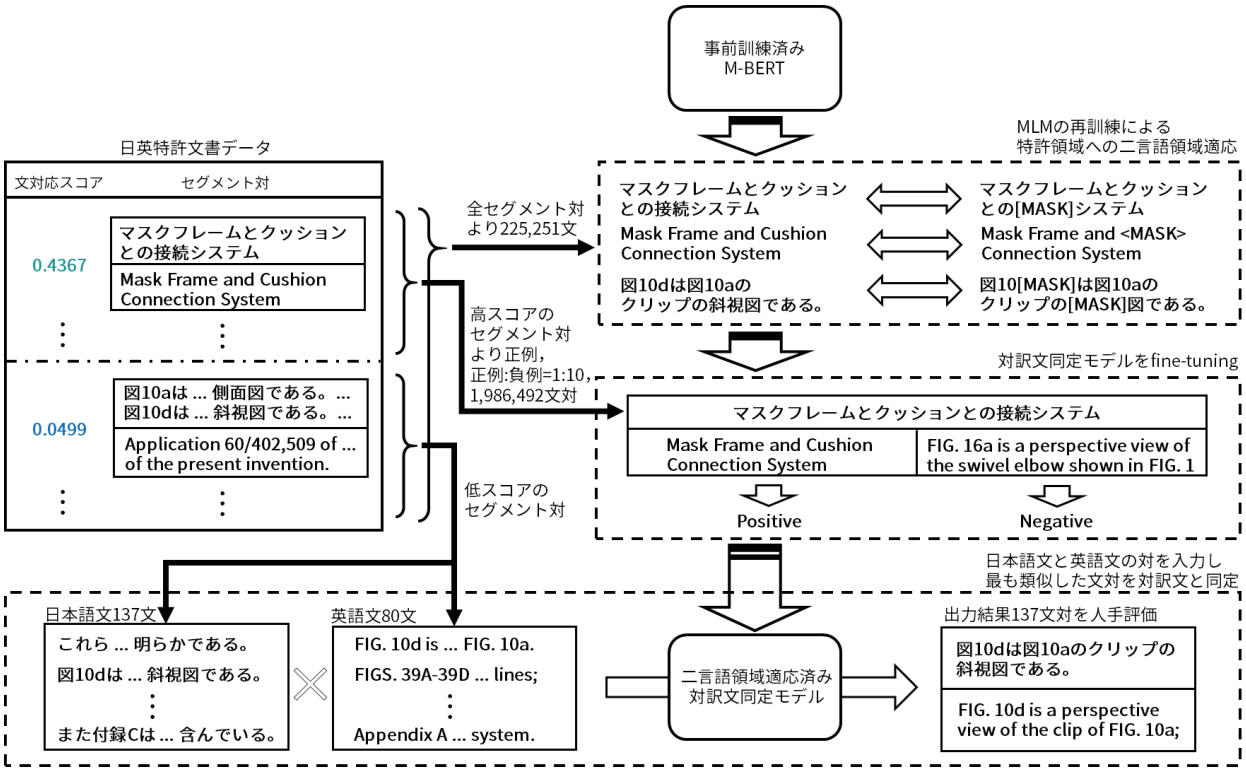

図 1: 対訳文同定の手順

fastBPE ツール [6] によるサブワードへの変換が行われた．また，文間距離の計算には Faiss [6] を用いた．

6 fine-tuning による対訳文分類器の訓練・評価

BERT [3] の評価にも用いられた GLUE ベンチマーク [14] の Quora Questoin Pairs (QQP) タスク [7] では，入力した二文が同一か否かの二値分類タスクを解く．一般に QQP タスクは単言語の二文に対して行うが，本論文では二言語領域適応した M-BERT を用い，異なる二言語の二文に対する二値分類タスクを解く．

正例には日英特許文書 240 文書のうち文対応スコアが 0.05 以上のセグメント対を文対単位に変換したものを利用した (図 1)．負例には正例の日本語文に対する LASER 文埋め込みの文間距離が近い英語文 10 文と，ランダムで取得した 10 文を組み合わせ (表 1)，正例と比例の比率が 1 対 10 になるようにデータを構築し，計 1,986,492 文対を得た (表 2(b))．また人手で作成した正例 60 件負例 40 件からなるデータを開発文に用いた．

fine-tuning では，1 エポックの訓練を行い [8]，その中で最も汎化誤差が小さいモデルを fine-tuning による対訳文同定モデルとして得た．

検証実験に用いる特許文書では 799 セグメント対を含む (表 3(a)) が，このうち文対応スコアが 0.05 未満のセグメント対 45 対を文対単位に変換したものを検証文集合とし，日本語文は 137 文，英語文は 80 文を得た (表 3(b))．

対訳文同定モデルの検証実験において，日本語文 137 文に対する英語文 80 文のすべての組み合わせにおける文類似度スコアを求め，文類似度スコアの最も高い一文を同定し，137 文対の出力結果を得た．その後人手評価によって，137 文対を適切な対訳と不適切な対訳に分類し，それぞれの数を集計した．

7 実験結果

表 4: 日英対訳特許文の同定結果の人手評価

手法	適切	不適切
内山法	3	134
LASER (ベースライン)	66	71
M-BERT + fine-tuning	82	55
M-BERT + 二言語領域適応 + fine-tuning	**88**	**49**

表 4 に対訳文同定タスクの実験結果を示す．LASER

[6] https://github.com/glample/fastBPE

[7] https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs

[8] NVIDIA Tesla P100 を二枚使用し，バッチサイズを 64 に設定したところ，1 エポックあたりの訓練時間は約 2 時間であった．

表 5: 対訳文同定結果の例

手法	文例	判定
日本語文	図 24e は、図 24a に示されているクッションの斜視図である。	–
内山法	Appendix B of incorporated U.S. Provisional Application of Moore et al., Ser. No. 60/402,509, includes pictures of two prior art masks discussed in the present invention.	不適切
LASER (ベースライン)	FIG. 24b is a frame side view of the cushion of FIG. 24a showing CAD construction lines;	不適切
M-BERT + fine-tuning	FIG. 25g is a cross-section taken along line 25g-25g of FIG. 25b;	不適切
M-BERT + 二言語領域適応 + fine-tuning	FIGS. 24c-24f illustrate various perspectives of the cushion shown in FIG. 24a;	適切

を使った場合には日本語文 137 文のうち 66 文が適切な対訳であったが，M-BERT に二言語領域適応と fine-tuning を行ったモデルでは日本語文 137 文のうち 88 文が適切な対訳であった．さらに fine-tuning のみを行ったモデルと比較した場合には 137 文のうち 82 文のみが適切な対訳であったため，二言語領域適応による効果が確認できた．また内山法による対訳文を評価すると，日本語文 137 文のうち 3 文しか適切な対訳が存在せず，文対応スコアが低いセグメント対から得た低品質の対訳文集合の中から一定の質の対訳文を得られる点で本論文の手法は有用であることが明らかとなった．

また，表 5 に実際の対訳文同定結果の例を示す．日本語文に対して，内山法による対訳文はまるで見当違いの文を指していることと，LASER や fine-tuning のみのモデルでは間違いではあるものの似ている文を出力できていること，二言語領域適応をしたモデルでは適切な文を出力できていることがわかる．これは特許領域の MLM が訓練されたことで図表番号のわずかな違いを認識できているためだと考えられる．

8 おわりに

本論文では，日英特許文書の領域において，二言語の非対訳特許文書から MLM を訓練することにより，M-BERT の文埋め込みを日英特許文書に対して二言語領域適応した．そして，二言語領域適応された Multilingual-BERT の文埋め込みを，さらに二値分類器として fine-tuning して，日英特許文書間の対訳文同定を行ったところ，領域適応しない手法よりも優れた同定結果を得た．今後の課題として，fine-tuning を行うことなく二言語領域適応のみで対訳文同定を行うことで，対訳文がない言語対においても対訳文同定タスクに取り組むことができるようになるため，非常に有用であると考えられる．

9 謝辞

本研究の実施において，日英特許文書データを提供していただいた，日本特許情報機構 (JAPIO) の関係者各位に深謝の意を表する．

参考文献

[1] M. Artetxe and H. Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *TACL*, pp. 597–610, 2019.

[2] V. Chaudhary, Y. Tang, F. Guzmán, H. Schwenk, and P. Koehn. Low-resource corpus filtering using multilingual sentence embeddings. In *Proc. 4th WMT*, pp. 263–268, 2019.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, pp. 4171–4186, 2019.

[4] M. Guo, Y. Yang Q. Shen and, H. Ge, D. Cer, G. H. Abrego, K. Stevens, N. Constant, Y. H. Sung, and R. Kurzweil B. Strope. Effective parallel corpus mining using bilingual sentence embeddings. In *Proc. 3rd WMT*, pp. 165–176, 2018.

[5] V. Hangya, F. Braune, A. Fraser, and H. Schutze. Two methods for domain adaptation of bilingual tasks: Delightfully simple and broadly applicable. In *Proc. 56th ACL*, pp. 810–820, 2018.

[6] J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017.

[7] M. Junczys-Dowmunt. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proc. 3rd WMT*, pp. 888–895, 2018.

[8] P. Koehn, F. Guzmán, V. Chaudhary, and J. Pino. Findings of the wmt 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proc. 4th WMT*, pp. 56–74, 2019.

[9] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL*, pp. 177–180, 2007.

[10] V. M. Sánchez-Cartagena, M. Banón, S. Ortiz-Rojas, and G. Ramírez-Sánchez. Prompsit's submission to wmt 2018 parallel corpus filtering shared task. In *Proc. 3rd WMT*, pp. 955–962, 2018.

[11] M. Uchiyama and H. Isahara. Reliable measures for aligning Japanese-English news articles and sentences. In *Proc. 41th ACL*, pp. 72–79, 2003.

[12] T. Utsuro, H. Ikeda, M. Yamane, Y. Matsumoto, and M. Nagao. Bilingual text, matching using bilingual dictionary and statistics. In *Proc. COLING*, pp. 1076–1082, 1994.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Proc. 30th NIPS*, pp. 5998–6008, 2017.

[14] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proc. EMNLP*, pp. 353–355, 2018.

[15] H. Xu and P. Koehn. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *Proc. EMNLP*, pp. 2945–2950, 2017.

[16] Y. Yang, G. H. Abrego, S. Yuan, M. Guo, Q. Shen, D. Cer, Y. H. Sung, B. Strope, and R. Kurzweil. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. *arXiv preprint arXiv:1902.08564.*, 2019.

[17] P. Zweigenbaum, S. Sharoff, and R. Rapp. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proc. 10th BUCC*, pp. 60–67, 2017.

[18] P. Zweigenbaum, S. Sharoff, and R. Rapp. Overview of the third BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proc. 11th BUCC*, pp. 39–42, 2018.