Artificial Intelligence In Medicine 153 (2024) 102889

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Oversampling effect in pretraining for bidirectional encoder representations from transformers (BERT) to localize medical BERT and enhance biomedical BERT

Shoya Wada[a,*], Toshihiro Takeda[a], Katsuki Okada[a], Shirou Manabe[a], Shozo Konishi[a], Jun Kamohara[b], Yasushi Matsumura[a]

[a] *Department of Medical Informatics, Osaka University Graduate School of Medicine, Japan*
[b] *Faculty of Medicine, Osaka University, Japan*

ARTICLE INFO

Keywords:
Natural language processing
Deep learning
BERT

ABSTRACT

Background: Pretraining large-scale neural language models on raw texts has made a significant contribution to improving transfer learning in natural language processing. With the introduction of transformer-based language models, such as bidirectional encoder representations from transformers (BERT), the performance of information extraction from free text has improved significantly in both the general and medical domains. However, it is difficult to train specific BERT models to perform well in domains for which few databases of a high quality and large size are publicly available.
Objective: We hypothesized that this problem could be addressed by oversampling a domain-specific corpus and using it for pretraining with a larger corpus in a balanced manner. In the present study, we verified our hypothesis by developing pretraining models using our method and evaluating their performance.
Methods: Our proposed method was based on the simultaneous pretraining of models with knowledge from distinct domains after oversampling. We conducted three experiments in which we generated (1) English biomedical BERT from a small biomedical corpus, (2) Japanese medical BERT from a small medical corpus, and (3) enhanced biomedical BERT pretrained with complete PubMed abstracts in a balanced manner. We then compared their performance with those of conventional models.
Results: Our English BERT pretrained using both general and small medical domain corpora performed sufficiently well for practical use on the biomedical language understanding evaluation (BLUE) benchmark. Moreover, our proposed method was more effective than the conventional methods for each biomedical corpus of the same corpus size in the general domain. Our Japanese medical BERT outperformed the other BERT models built using a conventional method for almost all the medical tasks. The model demonstrated the same trend as that of the first experiment in English. Further, our enhanced biomedical BERT model, which was not pretrained on clinical notes, achieved superior clinical and biomedical scores on the BLUE benchmark with an increase of 0.3 points in the clinical score and 0.5 points in the biomedical score. These scores were above those of the models trained without our proposed method.
Conclusions: Well-balanced pretraining using oversampling instances derived from a corpus appropriate for the target task allowed us to construct a high-performance BERT model.

1. Introduction

Pretraining large-scale neural language models on raw texts has been shown to improve the process of transfer learning considerably in natural language processing (NLP). With the introduction of transformer-based language models, such as bidirectional encoder representations from transformers (BERT), the information extraction performance of NLP from free text in the general domain has improved significantly [1,2]. Owing to the rapid increase in the volume of medical literature, the accuracy of information extraction in the biomedical domain is also expected to improve. Many studies have shown that pretraining BERT models on a large domain-specific text corpus, such as biomedical or

* Corresponding author at: Department of Medical Informatics, Osaka University Graduate School of Medicine, 2-2, Yamadaoka, Suita, Osaka 565-0871, Japan.
E-mail address: wada@hp-info.med.osaka-u.ac.jp (S. Wada).

https://doi.org/10.1016/j.artmed.2024.102889
Received 14 September 2022; Received in revised form 3 May 2024; Accepted 4 May 2024
Available online 5 May 2024

clinical texts, results in a satisfactory performance in their specific text-mining tasks [3–5]. Moreover, language models with an architecture other than BERT have been published in the medical field [6,7].

Nevertheless, significant barriers persist regarding the localization of medical BERT models. Publicly available high-quality, large-scale medical databases written in languages other than English that are sufficient to train BERT models are scant. For example, a subscription is required to perform a cross-search of Japanese medical journals, and most articles are published only in PDF format, which makes it difficult to obtain a large medical corpus. A high demand therefore exists for techniques that can build language models that work well even when the available resources are limited. In this regard, certain data augmentation techniques have been proposed for NLP [8]; however, no reports have been published on how oversampling affects the pretraining of BERT.

When the corpus used to develop a neural language model is small, the model is typically first pretrained on a large corpus and then trained on the small corpus for domain adaptation. However, this approach may introduce "catastrophic forgetting," which is the tendency of neural networks to fail to retain previously learned information completely after learning new information. No solution to this problem has yet been reported.

We hypothesized that the problem could be solved by oversampling a domain-specific corpus and using it for pretraining with a larger corpus in a balanced manner. Here, we describe our method and demonstrate that it can process an objective task with high performance. We propose the simultaneous pretraining of models with knowledge from different domains after oversampling. Accordingly, we also developed appropriate BERT models (Fig. 1). First, we applied our method to an English-based model and verified that the performance of the model was comparable to that of models built using existing methods. Second, we showed the improvement that our method offers over conventional models for medical tasks in Japanese. Third, we demonstrated that our approach enables the development of a pretrained model that enhances biomedical representation in both clinical and biomedical tasks by balancing the corpora used for pretraining.

2. Related works

2.1. Pretrained BERT variants

BERT [2] is a contextualized word representation model based on masked language modeling (MLM) that is pretrained using bidirectional transformers [1]. The BERT framework consists of two steps: pretraining and fine-tuning. During pretraining, the model is trained on an unlabeled large corpora. For fine-tuning, the BERT model is first initialized with pretrained weights, and all the weights are fine-tuned using labeled data from the downstream tasks.

The standard BERT model does not perform well in specialized domains, such as biomedical or scientific texts [3,9]. Two possible strategies could overcome this limitation: continual pretraining (CPT) on domain-specific corpora from an existing pretrained BERT model or pretraining from scratch (PTS) on domain-specific corpora [10]. The main benefit of CPT is that it leverages existing knowledge from a previously trained BERT model and thus requires fewer iterations to adapt the model to domain-specific nuances because the understanding of the foundational language has already been established. This efficiency results in a lower computational cost than PTS, which must learn both the basic structure of the language and the specialized knowledge of the domain from the ground up. The main advantage of PTS is the creation of a custom vocabulary tailored to the domain (see Section 3.1.2); however, the pretrained neural language model may be less adaptable if the number of documents in a specific domain is small.

Having introduced CPT and its importance for domain adaptation, it is essential to differentiate it from fine-tuning. While fine-tuning optimizes the pretrained model for specific tasks using labeled data, CPT extends the model's pretraining on domain-specific corpora. As a preparatory step, the CPT process enhances the model's familiarity with the domain's language and nuances. In this way, CPT makes the model more effective when it is subsequently fine-tuned for specific tasks within that domain.

Building upon the foundational concepts of CPT and PTS, many

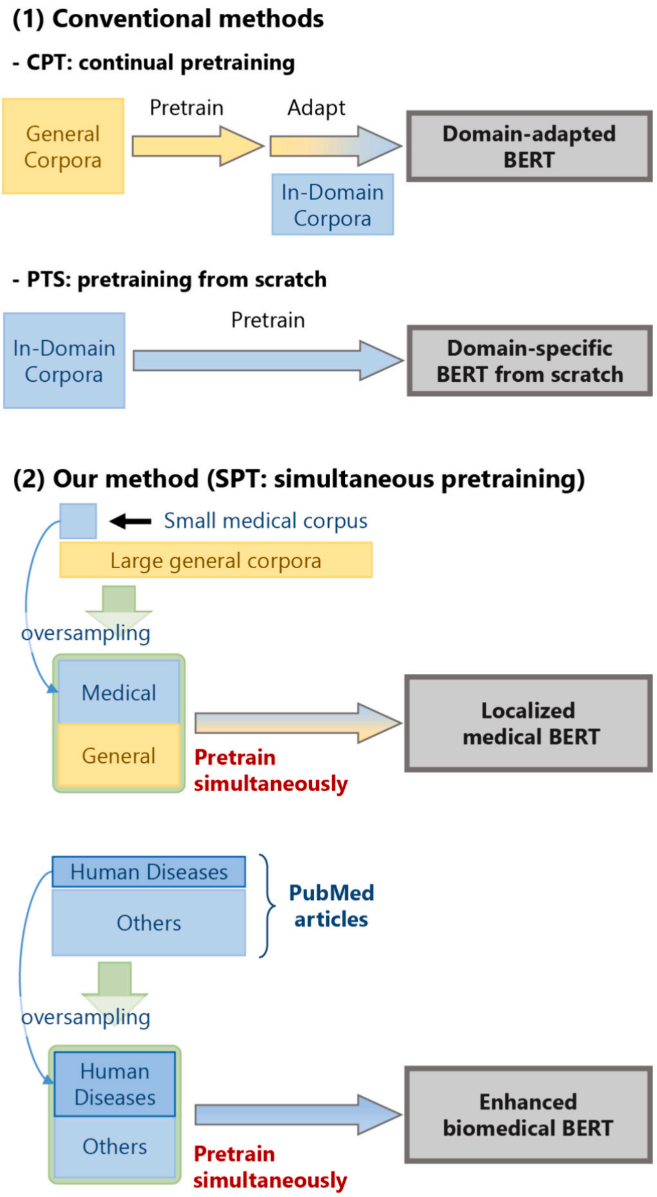

Fig. 1. Overview of pretraining BERT. Three approaches to pretraining BERT models are shown: (1) Conventional methods, which include continual pretraining (CPT), where general corpora are first pretrained, followed by domain adaptation using in-domain corpora to create a "domain-adapted BERT" model, and pretraining from scratch (PTS), where in-domain corpora are pretrained to construct a "domain-specific BERT from scratch" model. (2) Our simultaneous pretraining (SPT) method features two models: the "localized medical BERT" model (upper diagram), where the considerably smaller medical corpus is oversampled to match the size of the general corpora, followed by simultaneous pretraining of the model with knowledge from distinct domains. Similarly, the "enhanced biomedical BERT" model (lower diagram) distinguishes between and oversamples PubMed articles related to human diseases through metadata to match the size of the "other" corpus. This is followed by simultaneous pretraining, which also incorporates knowledge from the broader set of PubMed articles. Although not directly related to human diseases, this knowledge is still within the medical domain.

BERT variants have been developed, with each tailored to specific domains or tasks. These models leverage the flexibility of the original BERT architecture by combining it with the strengths of either CPT or PTS to enhance their domain-specific performance. In Table 1, we present a curated selection of publicly available BERT variants to exemplify this diversity and detail their pretraining strategies and focus domains.

BERT-Base was pretrained using the English Wikipedia (2500 million words) and BooksCorpus (800 million words) [2]. Some published models have been initialized from BERT-Base and trained using their domain-specific corpora. BioBERT is one of the first specialized versions of the BERT model and is tailored for the biomedical domain. It was pretrained using PubMed abstracts (4500 million words) and PubMed Central full-text articles (13,500 million words).

BlueBERT was published with the BLUE benchmark [11]. The authors evaluated BlueBERT-Base (P) and BlueBERT-Base (P + M), which were initialized from BERT-Base and additionally pretrained using only PubMed abstracts or a combination of PubMed abstracts for 5 million steps and MIMIC-III clinical notes [15] for 0.2 million steps, respectively. That is, their pretraining method is CPT. We refer to these as “biomedical BlueBERT” and “clinical BlueBERT,” respectively. The authors evaluated their models against biomedical and clinical tasks and found that clinical BlueBERT achieved the highest overall score on the combined tasks. However, in terms of the scores for the individual tasks within the benchmark, although the clinical model excelled in the clinical tasks, it performed less well in the biomedical tasks than biomedical BlueBERT. This discrepancy could have been due to catastrophic forgetting.

PubMedBERT is a notable domain-specific model that is pretrained using abstracts from PubMed and full-text articles from PubMed Central [4]. PubMedBERT incorporates several advanced pretraining techniques discovered by numerous researchers following the release of BERT, all of which have contributed to its effectiveness in biomedical NLP tasks. KeBioLM is a biomedical pretrained language model that leverages knowledge from the Unified Medical Language System (UMLS) knowledge bases [12]. The parameters of the transformers in KeBioLM are initialized from the checkpoint of PubMedBERT, and the vocabulary is derived from PubMedBERT. Specifically, KeBioLM extracts entities from PubMed abstracts, links them to UMLS, and applies text-entity fusion encoding. This approach has proven to be effective in tasks like named entity recognition (NER) and relation extraction.

A Japanese BERT model for the general domain was recently released by Tohoku University [13]. It was pretrained using the Japanese Wikipedia, and its vocabulary was obtained by applying byte-pair encoding (BPE) to the corpus. We refer to this model as the “Japanese Wiki (JWiki) model.”

In our proposed method, we aimed to adeptly utilize, from scratch, medical domain corpora that would traditionally be considered too small for pretraining language models to produce pretrained models with high robustness. Our approach differs from CPT and PTS. Further details are provided in Section 3.2.

Table 1
The publicly available models used in this paper.

Model	Pretraining method	Number of words	Domain
BERT-Base [2]	PTS	3300 M	General
BioBERT [3]	PTS	18,000 M	Biomedical
biomedical BlueBERT [11]	CPT from BERT-Base	>4000 M	Biomedical
clinical BlueBERT [11]	CPT from BERT-Base	>4500 M	Clinical
PubMedBERT [4]	PTS	16,800 M	Biomedical
KeBioLM [12]	CPT from PubMedBERT	N/S[†]	Biomedical
Japanese Wiki model [13]	PTS	550 M	General (Japanese)

CPT: continual pretraining. M: million. N/S: not specified. PTS: pretraining from scratch.

[†] The exact number of words in the corpus used for pretraining KeBioLM is not specified in the literature. However, it is noted that the utilized PubMedDS dataset comprises 3.5 million PubMed documents, which indicates a substantial biomedical text corpus. The precise word count remains undetermined. The Japanese corpora were tokenized using MeCab [14].

2.2. Oversampling techniques and their implications in NLP

Oversampling involves artificially augmenting the minority class in a dataset to match the size of the majority class and thereby create a more balanced distribution. This technique can be straightforward and entail, for example, duplicating instances of the minority class (simple oversampling), or more sophisticated, which may involve generating synthetic instances based on the characteristics of the minority class, such as the synthetic minority oversampling technique (SMOTE) and adaptive synthetic sampling (ADASYN) for numerical data [16,17].

In NLP, applying techniques such as SMOTE or ADASYN is difficult due to the inherent nature of textual data, which is fundamentally different from numerical data. Oversampling methods, while applicable, are often supplemented by text augmentation techniques to mitigate class imbalances. These text augmentation methods include synonym replacement, random insertion, random swap, and random deletion [8]. However, such approaches can introduce significant issues, particularly for downstream tasks. These synthetic texts may suffer from a loss of contextual and semantic information, which results in grammatically incorrect or structurally flawed text. Nonetheless, some studies have reported text augmentation as beneficial in specific fields, including the medical domain [8,18]. Recent studies have explored the use of language models for synthetic text generation as a form of oversampling [19]. This innovative approach leverages the fluent text generation capabilities of advanced language models and offers a promising solution to the problem of imbalanced data in NLP tasks.

For neural language models, oversampling is effective in fine-tuning BERT models with supervised data [20,21]. However, the impact of oversampling during the pretraining phase remains largely unexplored. In the present study, we proposed a novel approach that applies simple oversampling to corpora closely related to the target task domain during pretraining. The effectiveness of this method needed to be evaluated, with the expectation that it would not only maintain fundamental language comprehension abilities but also enhance the acquisition of domain-specific nuances.

3. Methods

Our models essentially have the same structure as BERT-Base [2]. We therefore begin this section with an overview of BERT. Next, we describe our method and refer to our models. Finally, we explain the fine-tuning process used to evaluate our models.

3.1. BERT

3.1.1. Pretraining

BERT pretraining is optimized for two unsupervised classification tasks (Fig. 2), the first of which is MLM. One training instance of MLM is a single modified sentence. Each token in the sentence has a 15 % chance of being selected for replacement. A selected token has an 80 % probability of being replaced by a special token [MASK], a 10 % probability of being replaced by another random token, or a 10 % probability of being left unmodified. The objective of MLM is a cross-entropy loss on predicting masked tokens.

The input instances consist of two text spans separated by a special token [SEP], where each span may include more than one sentence. A special instance marker [CLS] is added in front of each input example and used for next-sentence prediction (NSP). Tokens replaced by [MASK] are used for MLM.

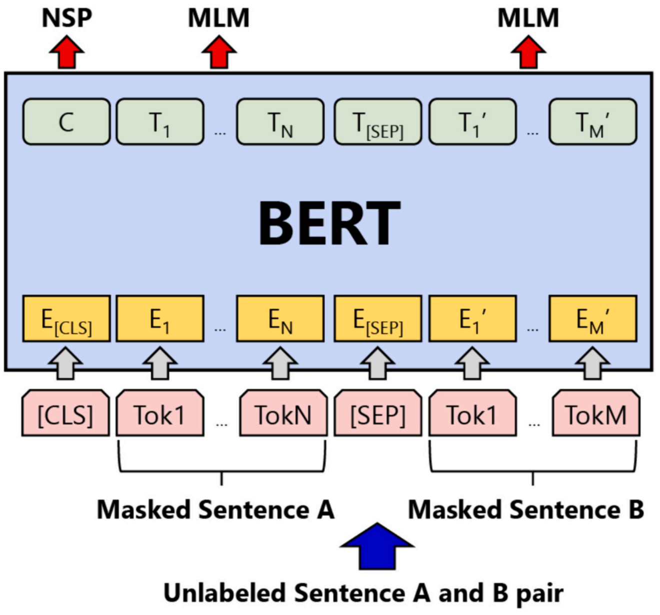

Fig. 2. Pretraining procedures for BERT (adapted from Devlin, Chang, Lee, and Toutanova [2]). The input instances consist of two sentences, such as text spans separated by a special token [SEP]. [CLS] is a special instance marker added in front of each input example and used for next-sentence prediction (NSP). The tokens replaced with [MASK] are used for masked language modeling (MLM).

The second unsupervised classification task is NSP, which is a binary classification loss for predicting whether two segments follow each other in the original text. Positive instances are created by taking consecutive sentences from the text corpus, and negative instances are created by pairing segments from different documents. Positive and negative instances are sampled with equal probabilities. NSP is designed to improve the performance of downstream tasks, such as natural language inference [22], which requires reasoning regarding the relationships between pairs of sentences.

3.1.2. *Vocabulary*

To manage the problem of out-of-vocabulary words, BERT uses vocabulary from subword units generated by WordPiece [23], which is based on BPE [24], for the unsupervised tokenization of the input text. The vocabulary is built such that it contains the most frequently used words or subword units. The main benefit of pretraining from scratch is to take advantage of a domain-specific custom vocabulary.

3.2. *Proposed method: Simultaneous pretraining*

It is generally known that if a BERT model is trained only on a small medical corpus, overfitting may degrade its performance. We hypothesized that this issue could be avoided if we simultaneously trained a BERT model using knowledge from both general and medical domains and that this could be achieved by simple oversampling. We therefore introduced a training method called simultaneous pretraining (SPT). Through this technique, we aimed to create pretraining instances efficiently from a set of corpora according to their file sizes and to pretrain a neural language model (Fig. 3). In the case of a medical BERT model, a small corpus corresponds to a medical corpus, and a large corpus is a general domain corpus, such as Wikipedia.

In the original implementation of BERT, oversampling is not possible because all the corpora are handled equally. However, in our method, the small and large corpora are first divided into smaller documents of the same target size (e.g., 10 MB), which may vary across different documents. This division process involves checking the disk size of each corpus and calculating the required number of divisions to reach the target size. Next, we obtain the line count for each document within the corpora and calculate the total line count. Using the previously calculated number of divisions, we calculate the approximate line count for the smaller documents and combine them so that the total line count falls within this range. Finally, while allowing for slight variations, we verify that the disk size of the divided documents is approximately in line with the target size and conclude the division process. When combining the documents, we ensure that the documents in the small and large corpora are of comparable file sizes with diverse combination patterns. In the present study, we chose to focus on file size as a language model-independent metric for corpus division rather than the total number of tokens.

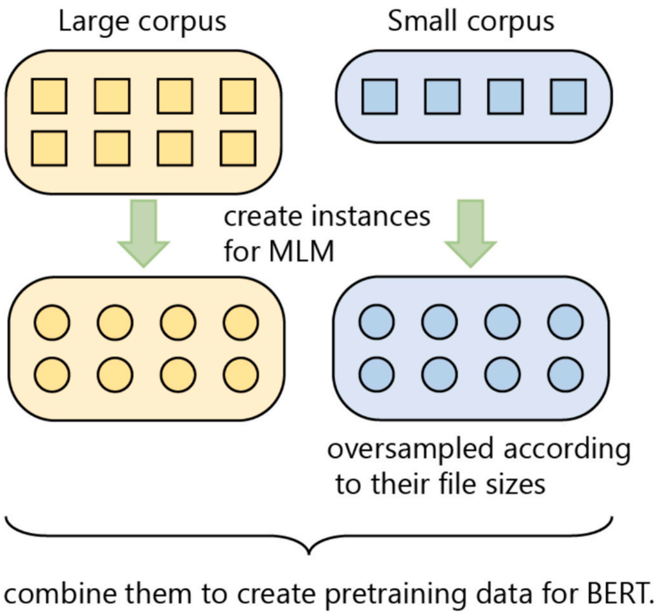

Fig. 3. The concept of our proposed method. The squares denote the raw texts in each corpus. The circles indicate the instances generated by processing them. When we generate instances for masked language modeling (MLM), we allow more instances to be generated from a small corpus than from a large corpus. We then combine all the instances in a balanced manner to create data for BERT pretraining.

For the oversampling process, we extract a fixed number from the sets generated by dividing the large corpus and then extract the same number from the small corpus's set without duplication. This extraction is randomized, but constraints are imposed to ensure that the probability of oversampling the documents is equally distributed across the corpus. By combining these two sets, we create an oversampling of the small corpus. As depicted in Fig. 3, documents from both the small and large corpora are combined to ensure equal proportions in terms of file size, thereby creating sufficient pretraining instances for the BERT model. After combining, the documents from the two corpora are mixed at a document level, and sentences may be combined as one instance only when generating instances where the NSP is false.

We introduce a balanced vocabulary that reflects terms that appear in the small corpus when we generate instances for pretraining. If we built a vocabulary with BPE without adjusting the file sizes of the small and large corpora, most of the words and subwords would be derived from the large corpus. To solve this problem, we duplicate random pieces of the small corpus until its file size is the same as that of the large corpus. This ensures that each training batch is randomly drawn with equal probability from either corpus. Subsequently, we construct the uncased vocabulary via BPE using tokenizers [25]. Balanced vocabulary learning is an option of SPT and is used when creating pretraining instances.

3.3. Our pretrained models and experimental settings

3.3.1. English biomedical BERT from a small biomedical corpus

To assess whether any disadvantages appeared during oversampling in BERT's pretraining, we evaluated our method in English. First, with a specific focus on biomedical literature rather than electronic health record documents, we empirically produced a limited corpus of clinically relevant articles from PubMed abstracts. PubMed comprises many citations for biomedical literature from MEDLINE; the database therefore constitutes a mix of articles from the fields of clinical medicine and life sciences. We first constructed a comprehensive medical corpus related to human diseases, which we denoted as the "clinical PubMed abstracts (clPMAs) corpus." This corpus was extracted from the PubMed abstracts using specific criteria for their medical subject heading (MeSH) IDs corresponding to unique tree numbers in the MeSH database. The selection was made based on abstracts containing MeSH IDs that began with "Diseases [C]" and explicitly excluded those with MeSH IDs corresponding to "Plant Structures [A18]," "Fungal Structures [A19]," "Bacterial Structures [A20]," "Viral Structures [A21]," "Archaea [B02]," and "Organism Forms [B05]." From the clPMAs corpus, we randomly extracted the target abstracts to create the "small PubMed abstracts (sPMAs) corpus," which aligned with the experimental conditions of the later-mentioned Japanese corpus. In addition to the sPMAs corpus, we created a general corpus randomly sampled from articles on the English Wikipedia, which were denoted as the "EWiki corpus." It was constructed to fit the later experiment in Japanese and was similar to the Japanese Wikipedia in terms of word count and file size. The combined methodological approach allowed us to focus the corpus on human diseases while avoiding unrelated structures and organisms.

The name given to the pretrained medical BERT model in English was SPT-sPMAs/EWiki. We used the sPMAs corpus as the small medical source and the EWiki corpus as the general corpus. The PTS-sPMAs and CPT-sPMAs<EWiki models were trained for comparison. The former was pretrained from scratch solely using the sPMAs corpus, and the latter was initialized from the EWiki model, which was pretrained using the EWiki corpus in our environment and trained using the sPMAs corpus for domain-specific adaptations similar to those of BioBERT [3].

Further, to validate the effectiveness of our SPT, we developed pretraining models using different sizes of small medical corpora and evaluated their performance. Specifically, we determined the size of each corpus using the file size of the sPMAs as a reference. We then selected four distinct corpus sizes, namely, 60 MB, 120 MB (to match the size of the sPMAs), 600 MB, and 1200 MB, and randomly selected abstracts from the clPMAs corpus. These sizes allowed us to assess how varying the corpus dimensions could affect the model's performance.

3.3.2. Japanese medical BERT from a small medical corpus

SPT-JCR/JWiki is the Japanese medical BERT model pretrained using our method. For the medical domain source, we used the medical reference titled "Today's Diagnosis and Treatment: Premium," which consists of 15 digital resources for clinicians in Japanese that have been published by Igaku-Shoin Co., Ltd. (referred to as the "Japanese clinical references" or "JCR"). Similarly, the JWiki corpus was used for the general domain.

Three pretrained BERT models were prepared for comparison purposes. One was a publicly available model, JWiki. The others were pretrained using conventional methods in our environment: CPT-JCR < JWiki, which was initialized with the JWiki model and trained for additional steps using Japanese clinical references, and PTS-JCR, which was pretrained from scratch using only Japanese clinical references.

3.3.3. Enhanced biomedical BERT from whole PubMed abstracts

Previous research has shown that domain-adaptive pretraining is effective and that additional task-adaptive pretraining enhances the performance of downstream tasks [26]. Although PubMed articles, which are commonly used in biomedical language models, constitute a mix of studies in the fields of clinical medicine and life sciences, biomedical NLP tasks are mainly focused on humans. We therefore hypothesized that our approach would boost the amount of training required on articles related to human diseases within the entire corpus of PubMed and evaluated this effect accordingly. We extracted a comprehensive set of PubMed abstracts related to human diseases to form a collection denoted as the "clinical PubMed abstracts (clPMAs) corpus," as detailed in Section 3.3.1. The other articles were referred to as other PubMed abstracts (oPMAs).

SPT-clPMAs/oPMAs is our enhanced biomedical BERT model, which was pretrained from scratch using all the PubMed abstracts. In this model, pretraining on medical articles, especially those related to human diseases, was emphasized through our method. For the training process, we utilized clPMAs as the small corpus and oPMAs as the large corpus. In addition to this model, we pretrained from scratch another one named PTS-ePMAs on the "entire PubMed abstracts (ePMAs) corpus" without oversampling. The hyperparameters were kept consistent between SPT-clPMAs/oPMAs and PTS-ePMAs, which allowed for a direct comparison between our SPT method and the PTS approach.

Table 2 lists the corpora used in our models, and Table 3 shows the pretrained models compared in this paper. For a comprehensive understanding of the acronyms and specific terminology used in this paper, refer to Table A1 in Appendix A.

4. Downstream tasks

Three evaluations were performed. First, we measured the BLUE benchmark scores of the SPT-sPMAs/EWiki model to demonstrate the effectiveness of our method in English. Second, we studied the performance of the Japanese medical BERT variants across four datasets to confirm that our method would be applicable in Japanese medical contexts. Third, we executed the BLUE benchmark with five different random seeds and compared the average score of the SPT-clPMAs/oPMAs model with those of PTS-ePMAs and some publicly available models to demonstrate the potential of our method.

4.1. BLUE benchmark

The BLUE benchmark, which comprises five different biomedical text-mining tasks with 10 corpora, was developed to facilitate research on language representations in the biomedical domain [11]. These 10 corpora are preexisting datasets widely used as shared tasks by the biomedical NLP community (Table 4). We chose the BLUE benchmark over the more recent BLURB leaderboard [4]. The tasks in the BLUE benchmark, including the clinical tasks, have been optimized for BlueBERT, with an input token length that rarely exceeds 128. This configuration allows for the construction and comparison of many pretrained

Table 2
List of the text corpora used for our models.

Abbr.	Corpus	Number of words	File size (GB)	Domain
JWiki	Japanese Wikipedia	550 M	2.6	(jp) General
JCR	Japanese clinical references	18 M	0.1	(jp) Medical
EWiki	Sampled English Wikipedia	500 M	3.0	(en) General
sPMAs	Small PubMed abstracts	18 M	0.1	(en) Medical
clPMAs	Clinical PubMed abstracts	280 M	1.8	(en) Medical
oPMAs	Other PubMed abstracts	2800 M	18	(en) Biomedical
ePMAs	Entire PubMed abstracts	3100 M	20	(en) Biomedical

en: English. jp: Japanese. M: million.
Notes: The Japanese corpora were tokenized using MeCab [14].

Table 3
List of the pretrained models evaluated in this paper.

	Model	Domain
Publicly available		
	BioBERT	(en) Biomedical
	biomedical BlueBERT	(en) Biomedical
	clinical BlueBERT	(en) Clinical
	PubMedBERT	(en) Biomedical
	KeBioLM	(en) Biomedical
	JWiki	(jp) General
Ours		
	PTS-EWiki	(en) General
	PTS-JCR	(jp) Medical
	PTS-sPMAs	(en) Biomedical
	PTS-ePMAs	(en) Biomedical
	CPT-JCR < JWiki	(jp) Medical
	CPT-sPMAs<EWiki	(en) Biomedical
	SPT-JCR/JWiki	(jp) Medical
	SPT-sPMAs/EWiki	(en) Biomedical
	SPT-clPMAs/oPMAs	(en) Biomedical

clPMAs: clinical PubMed abstracts. CPT: continual pretraining. ePMAs: entire PubMed abstracts. EWiki: sampled English Wikipedia. JCR: Japanese clinical references. JWiki: Japanese Wikipedia. oPMAs: other PubMed abstracts. PTS: pretraining from scratch. sPMAs: small PubMed abstracts. SPT: simultaneous pretraining.
Notes: The symbol "<" for CPT indicates the sequence of corpora used for pretraining. For example, "CPT-JCR < JWiki" signifies that the model was initially pretrained using the JWiki corpus, followed by additional training with the JCR corpus for domain adaptation. The symbol "/" for SPT denotes the simultaneous use of corpora for pretraining. For example, "SPT-JCR/JWiki" represents a model that was pretrained using both the JWiki corpus and an oversampled JCR corpus.

Table 4
BLUE tasks [11].

Corpus	Type	Task	Metrics	Domain
BC5CDR-disease	Mentions	Named-entity recognition	F1	Biomedical
BC5CDR-chemical	Mentions	Named-entity recognition	F1	Biomedical
BIOSSES	Sentence pairs	Sentence similarity	Pearson	Biomedical
ChemProt	Relations	Relation-extraction	Micro F1	Biomedical
DDI	Relations	Relation-extraction	Macro F1	Biomedical
HoC	Documents	Document classification	F1	Biomedical
i2b2 2010	Relations	Relation-extraction	Micro F1	Clinical
MedNLI	Pairs	Inference	Accuracy	Clinical
MedSTS	Sentence pairs	Sentence similarity	Pearson	Clinical
ShARe/CLEF	Mentions	Named-entity recognition	F1	Clinical

models and facilitates comprehensive evaluations encompassing various aspects of medical text mining. By prioritizing an input length of 128 instead of the 512 typically found in publicly available models, we were able to create and evaluate the models more quickly, which aligned with our research objectives.

Following the practice of Peng et al. [11], we used a macro-average of F1-scores and Pearson scores as a total score to compare the pretrained BERT models. Moreover, to evaluate the change in the total score achieved by our method, we calculated the scores of the clinical and biomedical domains individually as clinical and biomedical scores, respectively. The clinical score was the macro-average of MedSTS [27], ShARe/CLEF [28], i2b2 2010 [29], and MedNLI [30], and the biomedical score was the macro-average of BIOSSES [31], BC5CDR-disease/chemical [32], DDI [33], ChemProt [34], and the Hallmarks of Cancer (HoC) corpus [35]. In our study, the solution to each task followed the code published with the BLUE benchmark. For more information, refer to the original articles on the BLUE benchmark and each dataset.

4.2. Evaluation tasks in Japanese

NTCIR-13 MedWeb is a multilabel classification task in Japanese [36], which involves the classification of pseudo-tweets into eight labels that correspond to various diseases or symptoms. The evaluation metric used in our study was the micro-F1 score.

Although the NTCIR-13 MedWeb shared tasks involve classifying Japanese medical texts and consist of 2560 samples, they do not cater to medical textbooks—a crucial and unique corpus in our research. Accordingly, to broaden our validation, we also assessed our models against two specialized datasets, MedTxt-CR and MedTxt-RR, which were introduced at NTCIR-16 [37].

MedTxt-CR comprises 148 case reports from J-Stage open-access articles chosen explicitly for their clinical relevance, while MedTxt-RR consists of 135 radiology reports regarding lung cancer computed tomography. Both datasets, although smaller, are annotated with critical medical entities, which makes them appropriate for evaluating our models' entity recognition capabilities in the medical domain context. We evaluated them using the micro-F1 score.

While we recognized the value of the MedTxt-CR and MedTxt-RR datasets for specialized tasks, we remained mindful of their limited size. We therefore created a multiclass document classification task to address the lack of shared tasks specifically targeting medical domain documents derived from medical textbooks in Japanese. We based this task on the medical topics in the MSD Manual for the Professional [38] and named it DocClsJp. With a total of 2475 articles belonging to one of 22 disease categories, this task allowed us to utilize the first 128 tokens of each document as an input sentence and to define its disease category as the correct label. We employed five-fold stratified cross-validation to evaluate the results, with the micro-F1 score as the performance metric.

5. Experimental setup

To both pretrain BERT and fine-tune it for downstream tasks, we applied mixed-precision training called FP16 computation, which significantly accelerated the computation speed by performing operations in the half-precision format. We used two NVIDIA Quadro RTX 8000 (48 GB) graphics processing units (GPUs) for the pretraining and a single GPU for the fine-tuning.

5.1. Pretraining BERT

We modified the implementation released by NVIDIA to train our models [39], which enabled us to make use of FP16 computation, gradient accumulation, and a layer-wise adaptive moments-based (LAMB) optimizer [40]. Unless stated otherwise, the pretraining configuration was the same as that of BERT-Base.

5.1.1. Japanese medical BERT from a small medical corpus

For SPT-JCR/JWiki and PTS-JCR, the maximum sequence length was fixed at 128 tokens, and the global batch size (GBS) was set to 2048. Additionally, the LAMB optimizer was used with a learning rate (LR) of 7e–4. We trained the model for 125 K steps. The vocabulary size was 32 K. CPT-JCR < JWiki was initialized from the JWiki model and trained using the JCR corpus until the loss of MLM and NSP on the training dataset stopped decreasing. We also used the LAMB optimizer with an LR of 1e–4.

5.1.2. English biomedical corpus from a small biomedical corpus

We used the same settings for the SPT-sPMAs/EWiki, PTS-EWiki, and PTS-sPMAs models as for SPT-JCR/JWiki. CPT-sPMAs<EWiki was initialized from PTS-EWiki and trained using the sPMAs corpus with the same settings as for CPT-JCR < JWiki.

5.1.3. Enhanced biomedical BERT from whole PubMed abstracts

For SPT-clPMAs/oPMAs, we followed NVIDIA's implementation. We set the maximum sequence length to 128 tokens and trained the model for 7038 steps using a GBS of 65,536 and the LAMB optimizer with an LR of 6e–3. We then continued to train the model by allowing a sequence length of up to 512 tokens for an additional 1563 steps to learn positional embeddings using a GBS of 32,768 and the LAMB optimizer with an LR of 4e–3. The vocabulary size was 32 K. For PTS-ePMAs, we used the same settings as those applied in the SPT-clPMAs/oPMAs model without oversampling.

5.2. Fine-tuning BERT for downstream tasks

We mostly followed the same architecture and optimization provided in transformers for fine-tuning [25]. Additionally, we applied minimal architectural modifications to BERT's task-specific inputs and outputs and fine-tuned all the parameters in an end-to-end manner. Consequently, we set the maximum sequence length to 128 tokens in all settings and employed Adam [41] for fine-tuning using a batch size of 32 and an LR of 3e–5, 4e–5, or 5e–5. The number of training epochs was adjusted using a grid search. For each dataset and BERT variant, we selected the best LR and number of epochs on the development set and reported the corresponding test results.

We constructed models using multiple seed values in our fine-tuning process, which enabled robust evaluation. To examine the statistical properties of the scores, we applied the bootstrap method to reconstitute multiple evaluation datasets, with consideration of the mean values for analysis. Although some scores were analyzed for normality using the Shapiro–Wilk test, not all the scores had a normal distribution. To maintain consistency across all the experiments, we utilized the nonparametric Mann–Whitney *U* test to assess the significance of the differences between the models. A *p*-value <0.05 indicated a significant difference.

6. Results

Table 5 provides a summary of the performance of the SPT-sPMAs/EWiki model as measured by the BLUE score. For our proposed model, most of the individual scores in each dataset, with the exception of MedSTS, BIOSSES, DDI, ChemProt, and HoC, were statistically higher than those obtained using conventional methods. Consequently, our model outperformed the conventional models in these evaluations.

In Table 6, we present a comparison of the total scores of the BLUE benchmark with the same corpus size for EWiki and different biomedical corpus sizes. Both the CPT and PTS methods showed upward trends, even with the largest corpus size in our trial, but the growth was moderate. Our SPT method achieved statistically higher scores than these conventional methods for each biomedical corpus size examined, which indicates a significant difference in performance. Detailed individual scores corresponding to the different corpus sizes can be found in Table A2 in Appendix A.

A comparison of the micro-F1 scores of the models pretrained using our method across four datasets (DocClsJp, NTCIR-13 MedWeb, NTCIR-16 MedTxt-CR, and NTCIR-16 MedTxt-RR) is presented in Table 7. Our results for DocClsJp and NTCIR-16 MedTxt-RR demonstrated that the SPT-JCR/JWiki model had a statistically significantly higher performance over the other BERT-based models (constructed using either known or officially released techniques). For the NTCIR-16 MedTxt-CR dataset, the SPT model's performance was statistically significantly higher compared to the CPT-JCR < JWiki model, but no significant difference was evident in the comparison with the PTS-JCR model. Additionally, for the NTCIR-13 MedWeb dataset, although the CPT model showed the highest score, and the SPT model ranked second among the four models analyzed, the difference was not statistically significant.

Table 8 presents the summarized scores achieved by SPT-clPMAs/oPMAs on the BLUE benchmark compared with those of BioBERT, biomedical BlueBERT, clinical BlueBERT, PubMedBERT, KeBioLM, and

Table 5
BLUE scores of our BERT variants.

Model	Total	MedSTS	BIOSSES	BC5CDR-disease	BC5CDR-chemical	ShARe/CLEF	DDI	ChemProt	i2b2 2010	HoC	MedNLI
PTS-EWiki	79.2	83.1	89.2	82.3	90.8	75.2	76.2	64.6	69.4	84.7	76.3
PTS-sPMAs	80.5	**83.4**	**91.2**	84.0	91.1	76.9	77.6	**66.8**	71.4	**85.2**	77.0
CPT-sPMAs<EWiki	80.1	83.1	89.5	84.1	91.2	77.1	77.3	65.3	70.7	84.5	78.6
SPT-sPMAs/EWiki	**80.9 ± 0.1**[*]	83.1 ± 0.2	89.8 ± 0.5	**85.3 ± 0.1**[*]	**91.7 ± 0.1**[*]	**78.0 ± 0.1**[*]	**77.7 ± 0.2**	66.6 ± 0.1	**72.4 ± 0.1**[*]	84.8 ± 0.1	**79.5 ± 0.1**[*]

The bolded entries in the table indicate the model that achieved the highest score for each evaluation criterion (column).

[*] Indicates a significant difference ($p \leq 0.05$) in mean performance between the SPT-sPMAs/EWiki model and the other models.

Table 6
The effect of corpus size on each pretraining method based on the total scores of the BLUE benchmark.

Pretraining method	Biomedical corpus				
	9 M	18 M	94 M	190 M	(words)
	60 MB	120 MB	600 MB	1200 MB	(file size)
CPT	80.4	80.1	80.5	80.7	
PTS	80.2	80.5	80.6	81.1	
SPT	**80.8**$_{\pm 0.1}^{*}$	**80.9**$_{\pm 0.1}^{*}$	**81.4**$_{\pm 0.1}^{*}$	**82.0**$_{\pm 0.1}^{*}$	

CPT: continual pretraining from PTS-EWiki with each biomedical corpus. M: million. PTS: pretraining from scratch with EWiki and each biomedical corpus. SPT: simultaneous pretraining with EWiki and each biomedical corpus in a balanced manner.
The bolded entries in the table indicate the model that achieved the highest score for each evaluation criterion (column).
* Indicates a statistically significant difference ($p \leq 0.05$), thus demonstrating higher scores for the SPT method compared to the conventional methods across different biomedical corpus sizes.

Table 7
Test results for the DocClsJp, NTCIR-13 MedWeb, NTCIR-16 MedTxt-CR, and NTCIR-16 MedTxt-RR datasets.

BERT-based model	DocClsJp micro-F1	MedWeb micro-F1	MedTxt-CR micro-F1	MedTxt-RR micro-F1
JWiki	81.7	89.8	48.9	75.0
PTS-JCR	85.7	90.1	50.1	75.3
CPT-JCR < JWiki	85.2	**91.2**	48.0	74.3
SPT-JCR/JWiki	**87.2**$_{\pm 0.1}^{*}$	90.5$_{\pm 0.1}$	**50.2**$_{\pm 0.1}$	**76.2**$_{\pm 0.1}^{*}$

The bolded entries in the table indicate the model that achieved the highest score for each evaluation criterion (column).
* Indicates a significant difference ($p \leq 0.05$) in mean performance between the SPT-JCR/JWiki model and the other models.

PTS-ePMAs. PTS-ePMAs outperformed BioBERT and biomedical BlueBERT in the biomedical score category. Clinical BlueBERT, which had been initialized with biomedical BlueBERT and pretrained on MIMIC-III clinical notes, achieved the highest clinical score; however, its biomedical score was considerably lower than its initial score. PubMedBERT recorded a higher overall score. KeBioLM, which had been pretrained using PubMedBERT weights as the initial values and implemented with an architectural twist, surpassed the accuracy of PubMedBERT in a few tasks, and its clinical score also increased slightly. However, the total and biomedical scores, which were primarily influenced by a significantly lower BIOSSES score of 68.5, decreased. In contrast, the biomedical score of our SPT-clPMAs/oPMAs model did not decrease. Notably, its clinical score increased compared with that of the PTS-ePMAs model. A comparison of these two models revealed that our SPT model scored significantly higher than the PTS-ePMAs model, except in the BC5CDR-chemical, DDI, and MedSTS tasks.

Table 9 illustrates the performance differences between the various models across the biomedical, clinical, and total scores. In column (A), using clinical notes in pretraining led to a more significant clinical score increase of 1.2 than in the other comparisons. Column (B) explores the architectural changes and external knowledge use and shows a slight but non-significant clinical score increase of 0.16 ($p = 0.064$). In column (C), the comparison between the conventional approach and our proposed method reveals significant differences in the biomedical and clinical scores, with our models outperforming the other combinations. These comparisons underscore the effectiveness of our approach and highlight the particular gains in clinical adaptation and balanced performance across the domains.

7. Discussion

7.1. Principal results

The models trained using our method proved robust on the BLUE benchmark even when using a small medical corpus. Furthermore, we demonstrated that our method could be used to construct both localized medical BERT and enhanced biomedical BERT. These results highlight the importance of adapting the corpus used for pretraining to the target task as well as the effectiveness of our proposed method in supporting this.

We first created our SPT-sPMAs/EWiki model by combining a small biomedical corpus and a large general corpus in English. This model performed sufficiently well in practice. The PTS-sPMAs model

Table 9
Comparative analysis of the models based on the biomedical, clinical, and total scores.

	(A) BlueBERT: biomedical -> clinical	(B) PubMedBERT -> KeBioLM	(C) Ours: PTS-ePMAs -> SPT-clPMAs/oPMAs
Biomedical score	−2.88	−4.01	**+0.49**$_{\pm 0.06}^{**}$
Clinical score	**+1.15**	+0.16	+0.34$_{\pm 0.06}^{*}$
Total score	−1.27	−2.34	**+0.43**$_{\pm 0.05}^{**}$

The bolded entries in the table indicate the model that achieved the highest score for each evaluation criterion (row).
* $p \leq 0.05$, which denotes a significant difference, where the mean performance of (C) is better than (B).
** $p \leq 0.05$, which indicates a significant difference, where the mean performance of (C) is better than that of both (A) and (B).

Table 8
Performance of our models for the tasks in the BLUE benchmark.

Task	Bio-BERT	biomedical BlueBERT	clinical BlueBERT	PubMed-BERT	KeBio-LM	PTS-ePMAs	SPT-clPMAs/oPMAs
BIOSSES	89.2	89.6	81.7	92.8	68.5	92.5	**93.4**$_{\pm 0.3}^{*}$
BC5CDR-disease	85.8	86.3	85.0	86.8	**87.5**	86.6	87.3$_{\pm 0.1}^{*}$
BC5CDR-chemical	93.3	93.4	92.4	93.9	93.7	**94.1**	94.0$_{\pm 0.1}$
DDI	79.1	79.4	78.7	79.9	**80.2**	79.6	79.5$_{\pm 0.2}$
ChemProt	73.3	73.5	68.5	74.5	74.3	74.1	**75.4**$_{\pm 0.1}^{*}$
HoC	85.8	86.4	85.1	**86.6**	86.2	85.9	86.2$_{\pm 0.1}^{*}$
ShARe/CLEF	78.3	78.6	**80.1**	80.0	79.5	79.1	79.9$_{\pm 0.1}^{*}$
MedSTS	84.7	84.6	84.2	**85.0**	84.8	84.7	84.2$_{\pm 0.2}$
i2b2 2010	74.4	73.9	**75.7**	74.6	74.9	74.1	74.7$_{\pm 0.1}^{*}$
MedNLI	83.2	82.1	83.9	83.2	**84.2**	82.5	83.0$_{\pm 0.1}^{*}$
Biomedical score	84.4	84.8	81.9	85.7	81.7	85.5	**86.0**$_{\pm 0.1}^{*}$
Clinical score	80.1	79.8	**81.0**	80.7	80.8	80.1	80.4$_{\pm 0.1}^{*}$
Total score	82.7	82.8	81.5	83.7	81.4	83.3	**83.8**$_{\pm 0.1}^{*}$

The bolded entries in the table indicate the model that achieved the highest score for each evaluation criterion (row).
* $p \leq 0.05$, which denotes that the mean performance of SPT-clPMAs/oPMAs was statistically significantly higher than that of PTS-ePMAs.

performed worse than the PTS-EWiki model, and the CPT-sPMAs<EWiki model had a higher score than PTS-EWiki. In contrast, our SPT-sPMAs/EWiki model achieved a superior score compared to the others. This result suggests that the model pretrained via conventional methods suffered from overfitting and that our method can avoid this issue.

Our test of different biomedical corpus sizes additionally revealed the advantages of our method. The performance of both the CPT and PTS models increased with the biomedical corpus volume, albeit gradually. In comparison, our method was superior to the others for every size of the biomedical corpus. This result supported the effectiveness of our method using a small corpus in English; it therefore had the potential to also be applied to other languages.

We applied our method to medical BERT in Japanese for four tasks. In DocClsJp, our SPT-JCR/JWiki model outperformed the other BERT variants. PTS-JCR performed better than the JWiki model and was comparable to the CPT model, while our SPT model scored 2.0 points higher in the micro-F1 score. This result confirmed the versatility of our method, which was consistent with the first experiment.

Building on the success observed with the DocClsJp task, we extended our evaluation to NER tasks, where we utilized the NTCIR-16 MedTxt-CR and MedTxt-RR datasets. The SPT-JCR/JWiki model consistently demonstrated superior performance, which underscored its robustness across diverse tasks. Notably, within the MedTxt-RR dataset, the SPT-JCR/JWiki model achieved statistically significant results, thus reinforcing the efficacy of our approach. At the same time, the performance gap between the SPT-JCR/JWiki and PTS-JCR models with the MedTxt-CR dataset was not statistically significant, which could potentially be attributed to the smaller size of the dataset but nonetheless offered valuable insights. Specifically, the CPT-JCR < JWiki model's underperformance among the four models tested suggested that the absence of a task-aligned vocabulary hampered its effectiveness in the NER tasks, as discussed in Section 3.1.2. Moreover, this outcome implied that utilizing a small medical corpus for CPT may compromise general language processing capabilities, which merited further investigation.

However, NTCIR-13 MedWeb showed no statistically significant difference between the SPT and CPT models. This discrepancy may have been due to the NTCIR-13 MedWeb text's composition of pseudo-tweets, which mimic general public expression in a colloquial style. This informal style may not have aligned well with our method, which emphasizes medical textbook knowledge during pretraining and thereby explains the lack of a significant difference in the results.

Notably, we found that a high-performance pretrained model can be trained using our method with the SPT-clPMAs/oPMAs model. The results of the PTS-ePMAs model demonstrated that the configuration we used in the pretraining of the BERT models was the most significant factor responsible for the improvement in their scores compared with BioBERT, biomedical BlueBERT, and PubMedBERT. Previous studies have reported that larger batch sizes and longer steps in pretraining are effective in improving performance [40,42]; PTS-ePMAs, SPT-clPMAs/oPMAs, PubMedBERT, and KeBioLM therefore likely benefited from these features. Furthermore, our simultaneous pretraining with oversampling achieved an improvement in the BLUE benchmark scores, especially in the clinical scores, although we used only PubMed abstracts rather than clinical notes. The results for clinical BlueBERT showed that the corpus used for the last pretraining had the greatest effect on performance but that its high clinical score came at the expense of the biomedical score. The success of our model showed that enhanced pretraining of biomedical articles close to a specific task can improve both clinical and biomedical scores at the same time, even when the available language resources are limited.

The common factor across the three experiments was the scarcity of suitable resources available for pretraining for the target tasks. It was clear that pretraining on corpora corresponding to target tasks would be necessary; however, no solution to the lack of such corpora was evident. While the release of PubMedBERT raised some questions about the need to address both clinical and biomedical tasks [4], we demonstrated that our oversampling method could solve this problem. Additionally, unlike KeBioLM, we did not modify the architecture of BERT in our method, as it was highly advantageous to be able to use the existing code.

7.2. Limitations

The present study had several notable limitations. First, we checked the robustness of our models on multiple tasks in English; however, we evaluated the SPT-JCR/JWiki model for only four tasks in Japanese. This was because few text-mining shared tasks were available in Japanese for highly specialized medical documents [43], and it was difficult to solve this problem directly. Japanese documents differ from those in English in that words are not separated by spaces, and tokenization for Japanese documents therefore requires a morphological analysis. Nevertheless, we applied a common process in the present study. These verifications indicate that our method could have general applicability. Second, we focused on the language models of the BERT architecture. The effectiveness of our method with respect to other architectures is thus unknown [10]. Additionally, the main goal of our study was not to construct a "state-of-the-art" model but to assess the generality of our proposed method. To this end, we needed to build and compare numerous models quickly. Accordingly, we adopted the BLUE benchmark with a token count of 128, which was sufficient for our evaluation. However, the choice of this specific sequence length may seem short when applied to long clinical documents, and its impact on the results warrants further investigation. Third, we explored only one-to-one balancing in the oversampling process. The exploration of other ratios of general-to-biomedical data may yield different results and could be the subject of future research. Fourth, in the case of BIOSSES, the performance of KeBioLM was significantly lower, which prevented us from using the biomedical score for comparison. BIOSSES has a notably small number of training data compared to other shared tasks (train:development:test = 64:16:20), and this leads to the construction of unstable models [11]. Nonetheless, our constructed models produced stable and high approximations within BIOSSES, which may indicate our model's robustness. Fifth, exploring well-balanced pretraining with PubMed and MIMIC-III datasets via our SPT method presents a significant avenue for future research, as such investigations could potentially optimize pretraining strategies for medical language models beyond the sequential approach used in clinical BlueBERT. This area, which was not explored in our present study due to its scope, promises to refine the performance and robustness of NLP models in the medical domain.

8. Conclusion

In this study, we introduced a pretraining technique based on oversampling and produced high-performance BERT models that could deal with targeted tasks. First, we showed that a practical medical BERT model could be constructed via our method using a small medical corpus in English and that it could then be applied in Japanese. Second, we confirmed that a pretrained biomedical model that also manages clinical tasks could be produced using our method. These results support the validity of our hypotheses. Our study could thus help overcome the challenges of biomedical text-mining tasks in both English and other languages.

Funding

This work was supported by the Council for Science, Technology and Innovation, Cross-ministerial Strategic Innovation Promotion Program, "Innovative AI Hospital System" (Funding Agency: National Institute of Biomedical Innovation, Health, and Nutrition). The funding body had no role in the design of the study; collection, analysis, and interpretation of data; or in writing the manuscript.

CRediT authorship contribution statement

Shoya Wada: Writing – original draft, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Toshihiro Takeda:** Writing – review & editing, Supervision. **Katsuki Okada:** Writing – review & editing. **Shirou Manabe:** Writing – review & editing. **Shozo Konishi:** Writing – review & editing. **Jun Kamohara:** Validation, Investigation. **Yasushi Matsumura:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of Generative AI and AI-assisted technologies in the writing process

statement: During the preparation of this work, the authors used DeepL (www.deepl.com) to assist in translating the original text into English, Grammarly (grammarly.com) to check for grammatical errors and confirm alternative expressions, and ChatGPT (chat.openai.com) to aid in expressing the intended message in English, perform grammar checks, and insert new descriptions in a natural expression. After using these tools/services, the authors reviewed and edited the content as needed; therefore, they take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no conflict of interests.

Data availability

Our SPT-clPMAs/oPMAs model, an enhanced biomedical BERT model, and the source code for fine-tuning are available freely on the models' hub of Hugging Face and GitHub [44,45]. The pretrained weights of Japanese medical BERT models in this study are available for academic purposes on GitHub [46]. The DocClsJp dataset is not publicly available because of restrictions on the secondary distribution of copyrighted works, but it is available from the corresponding author upon reasonable request.

Appendix A

Table A1
Glossary of acronyms and terms - a detailed list of the specific acronyms, abbreviations, and unique terms used throughout the paper, providing definitions and context for enhanced understanding.

Abbreviation	Description
Pretraining Method	
CPT	**Continual pretraining**: An approach that takes pretrained model parameters as initial values, and conducts additional pretraining on a targeted domain corpus for domain adaptation.
PTS	**Pretraining from scratch**: A method that initializes the model's parameters with random values and begins pretraining from that.
SPT	**Simultaneous pretraining**: Our proposed technique for creating pretraining instances efficiently and pretraining a neural language model. Further details are described in Section 3.2.
Corpus	
clPMAs	**Clinical PubMed abstracts**: A biomedical corpus related to human diseases, extracted from PubMed abstracts using specific criteria described in Section 3.3.1.
ePMAs	**Entire PubMed abstracts**: A name for the PTS because this method does not discriminate between corpus types.
EWiki	**English Wikipedia corpus**: A general corpus randomly sampled from articles on English Wikipedia, constructed to fit the experiment in Japanese.
JCR	**Japanese clinical references**.
JWiki	**Japanese Wikipedia corpus**.
oPMAs	**Other PubMed abstracts**: PubMed abstracts other than the clPMAs corpus.
sPMAs	**Small PubMed abstracts**: A randomly extracted corpus from the clPMAs corpus, aligning with the experimental conditions.
Model	
CPT-JCR < JWiki	Pretrained first using the JWiki corpus, and then the JCR corpus.
CPT-sPMAs<EWiki	Pretrained first using the EWiki corpus, and then the sPMAs corpus.
PTS-ePMAs	Pretrained with PTS, using the ePMAs corpus.
PTS-EWiki	Pretrained with PTS, using the EWiki corpus.
PTS-JCR	Pretrained with PTS, using only the JCR corpus.
PTS-sPMAs	Pretrained with PTS, using only the sPMAs corpus.
SPT-clPMAs/oPMAs	Pretrained with SPT using the oPMAs corpus and the oversampled clPMAs corpus.
SPT-JCR/JWiki	Pretrained with SPT using the JWiki corpus and the oversampled JCR corpus.
SPT-sPMAs/EWiki	Pretrained with SPT using the EWiki corpus and the oversampled sPMAs corpus.

The symbol "<" indicates the sequence of corpora used for Continual Pretraining (CPT). The symbol "/" denotes the simultaneous use of corpora for Simultaneous Pretraining (SPT).

Table A2
Further detailed individual scores corresponding to different corpus sizes can be found in Appendix Table 2 (Table A2).

	CPT				PTS				SPT			
Task	60 MB	120 MB	600 MB	1200 MB	60 MB	120 MB	600 MB	1200 MB	60 MB	120 MB	600 MB	1200 MB
BIOSSES	90.8	89.5	87.9	89.3	88.8	91.2	86.1	87.3	91.7	89.8	88.3	**92.2**
BC5CDR-disease	84.4	84.1	84.6	84.5	83.7	84.0	84.8	85.2	84.9	85.3	85.8	**86.1**
BC5CDR-chemical	91.2	91.2	91.5	91.5	91.0	91.1	91.8	92.3	91.3	91.7	92.8	**93.1**
DDI	77.4	77.3	78.0	78.3	77.9	77.6	77.4	78.2	**78.4**	77.7	77.8	78.3

(*continued on next page*)

Table A2 (*continued*)

	CPT				PTS				SPT			
Task	60 MB	120 MB	600 MB	1200 MB	60 MB	120 MB	600 MB	1200 MB	60 MB	120 MB	600 MB	1200 MB
ChemProt	65.1	65.3	66.8	67.3	67.0	66.8	67.6	69.5	65.8	66.6	70.7	**70.7**
HoC	85.7	84.5	85.4	85.1	84.4	85.2	86.1	85.7	85.2	84.8	**86.1**	85.5
ShARe/CLEF	77.3	77.1	77.1	77.2	76.1	76.9	77.0	77.4	77.5	78.0	**78.1**	78.1
MedSTS	82.9	83.1	83.3	83.2	83.6	83.4	83.5	83.5	82.3	83.1	83.0	**83.8**
i2b2 2010	71.3	70.7	71.3	70.8	71.2	71.4	72.0	71.4	71.7	**72.4**	71.7	71.7
MedNLI	78.2	78.6	79.1	79.8	78.2	77.0	79.4	80.3	79.1	79.5	80.4	**80.7**
Biomedical Score	82.4	82.0	82.4	82.7	82.1	82.7	82.3	83.0	82.9	82.7	83.6	**84.3**
Clinical Score	77.4	77.4	77.7	77.7	77.2	77.2	78.0	78.1	77.7	78.2	78.3	**78.6**
Total Score	80.4	80.1	80.5	80.7	80.2	80.5	80.6	81.1	80.8	80.9	81.4	**82.0**

CPT: continual pretraining from PTS-EWiki with each biomedical corpus. PTS: pretraining from scratch with EWiki and each biomedical corpus. SPT: simultaneous pretraining with EWiki and each biomedical corpus in a balanced manner.

The bolded entries in the table indicate the model that achieved the highest score for each evaluation criterion (row).

Notes: Column names correspond to the disk size of the small PubMed abstracts (sPMAs) corpus used during the construction of each pre-training model.

References

[1] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst 2017;30.

[2] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers); 2019. https://doi.org/10.18653/v1/N19-1423.

[3] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36:1234–40. https://doi.org/10.1093/bioinformatics/btz682.

[4] Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare 2022;3:1–23. https://doi.org/10.1145/3458754.

[5] Alsentzer E, Murphy J, Boag W, Weng W-H, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd clinical natural language processing workshop; 2019. https://doi.org/10.18653/v1/w19-1909.

[6] Shin H-C, Zhang Y, Bakhturina E, Puri R, Patwary M, Shoeybi M, et al. BioMegatron: larger biomedical domain language model. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP); 2020. https://doi.org/10.18653/v1/2020.emnlp-main.379.

[7] Naseem U, Dunn AG, Khushi M, Kim J. Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT. BMC Bioinformatics 2022;23:144. https://doi.org/10.1186/s12859-022-04688-w.

[8] Wei J, Zou K. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP); 2019. https://doi.org/10.18653/v1/d19-1670.

[9] Beltagy I, Lo K, Cohan A. 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 (emnlp-ijcnLP); 2019. https://doi.org/10.18653/v1/d19-1371.

[10] Kalyan KS, Rajasekharan A, Sangeetha S. AMMU: a survey of transformer-based biomedical pretrained language models. J Biomed Inform 2022;126:103982. https://doi.org/10.1016/j.jbi.2021.103982.

[11] Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In: Proceedings of the 18th BioNLP workshop and shared task; 2019. https://doi.org/10.18653/v1/w19-5006.

[12] Yuan Z, Liu Y, Tan C, Huang S, Huang F. Improving biomedical pretrained language models with knowledge. In: Proceedings of the 20th workshop on biomedical language processing; 2021. p. 180–90.

[13] Suzuki M. cl-tohoku/bert-japanese: BERT models for Japanese text. https://github.com/cl-tohoku/bert-japanese. [Accessed 13 January 2020].

[14] Kudo T, Yamamoto K, Matsumoto Y. Applying conditional random fields to Japanese morphological analysis. In: Proceedings of the 2004 conference on empirical methods in natural language processing; 2004.

[15] Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific data 2016;3:160035. https://doi.org/10.1038/sdata.2016.35.

[16] He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence): Ieee; 2008. p. 1322–8.

[17] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57. https://doi.org/10.1613/jair.953.

[18] Ishikawa T, Yakoh T, Urushihara H. An NLP-inspired data augmentation method for adverse event prediction using an imbalanced healthcare dataset. IEEE Access 2022;10:81166–76. https://doi.org/10.1109/ACCESS.2022.3195212.

[19] Shaikh S, Daudpota SM, Imran AS, Kastrati Z. Towards improved classification accuracy on highly imbalanced text dataset using deep neural language models. Appl Sci 2021;11:869. https://doi.org/10.3390/app11020869.

[20] Yoosuf S, Yang Y. Fine-grained propaganda detection with fine-tuned BERT. In: Proceedings of the second workshop on natural language processing for internet freedom: censorship, disinformation, and propaganda; 2019. p. 87–91.

[21] Shi Y, ValizadehAslani T, Wang J, Ren P, Zhang Y, Hu M, et al. Improving imbalanced learning by pre-finetuning with data augmentation. In: Fourth international workshop on learning with imbalanced domains: theory and applications: PMLR; 2022. p. 68–82.

[22] Bowman S, Angeli G, Potts C, Manning CD. A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 conference on empirical methods in natural language processing; 2015. https://doi.org/10.18653/v1/d15-1075.

[23] Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, et al. Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint arXiv:160908144. 2016. https://doi.org/10.48550/arXiv.1609.08144.

[24] Sennrich R, Haddow B, Birch A. 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. https://doi.org/10.18653/v1/p16-1162.

[25] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations; 2020. https://doi.org/10.18653/v1/2020.emnlp-demos.6.

[26] Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, et al. Don't stop pretraining: adapt language models to domains and tasks. In: Proceedings of the 58th annual meeting of the association for computational linguistics; 2020. https://doi.org/10.18653/v1/2020.acl-main.740.

[27] Wang Y, Afzal N, Fu S, Wang L, Shen F, Rastegar-Mojarad M, et al. MedSTS: a resource for clinical semantic textual similarity. Lang Resour Eval 2020;54:57–72. https://doi.org/10.1007/s10579-018-9431-1.

[28] Suominen H, Salanterä S, Velupillai S, Chapman WW, Savova G, Elhadad N, et al. Overview of the ShARe/CLEF eHealth evaluation lab 2013. International conference of the cross-language evaluation forum for european languages. 2013. https://doi.org/10.1007/978-3-642-40802-1_24.

[29] Uzuner Ö, South BR, Shen S, Duvall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011;18:552–6. https://doi.org/10.1136/amiajnl-2011-000203.

[30] Romanov A, Shivade C. Lessons from natural language inference in the clinical domain. In: Proceedings of the 2018 conference on empirical methods in natural language processing; 2018. https://doi.org/10.18653/v1/d18-1187.

[31] Soğancıoğlu G, Öztürk H, Özgür A. BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. Bioinformatics 2017;33:i49–58. https://doi.org/10.1093/bioinformatics/btx238.

[32] Li J, Sun Y, Johnson RJ, Sciaky D, Wei C-H, Leaman R, et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database 2016: baw068. https://doi.org/10.1093/database/baw068.

[33] Herrero-Zazo M, Segura-Bedmar I, Martínez P, Declerck T. The DDI corpus: an annotated corpus with pharmacological substances and drug–drug interactions. J Biomed Inform 2013;46:914–20. https://doi.org/10.1016/j.jbi.2013.07.011.

[34] Krallinger M, Rabal O, Akhondi SA. Overview of the BioCreative VI chemical-protein interaction Track. In: Proceedings of the sixth BioCreative challenge evaluation workshop; 2017. p. 141–6.

[35] Baker S, Silins I, Guo Y, Ali I, Högberg J, Stenius U, et al. Automatic semantic classification of scientific literature according to the hallmarks of cancer. Bioinformatics 2016;32:432–40. https://doi.org/10.1093/bioinformatics/btv585.

[36] Wakamiya S, Morita M, Kano Y, Ohkuma T, Aramaki E. Overview of the NTCIR-13: MedWeb Task. In: The 13th NTCIR conference on evaluation of information access technologies; 2017. p. 40–9.

[37] Yada S, Nakamura Y, Wakamiya S, Aramaki E. Real-MedNLP: overview of REAL document-based MEDical natural language processing task. In: Proceedings of the 16th NTCIR conference on evaluation of information access technologies (NTCIR-16); 2022.

[38] Merck. MSD manual for the professional. https://www.msdmanuals.com/ja-jp/professional. [Accessed 3 December 2019].

[39] NVIDIA. BERT for PyTorch. https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/LanguageModeling/BERT. [Accessed 24 January 2020].

[40] You Y, Li J, Reddi S, Hseu J, Kumar S, Bhojanapalli S, et al. Large batch optimization for deep learning: training bert in 76 minutes. In: International Conference on Learning Representations; 2020. https://doi.org/10.48550/arXiv.1904.00962.

[41] Kingma DP, Ba J. Adam: a method for stochastic optimization. In: Proceedings of international conference on learning representations; 2015. https://doi.org/10.48550/arXiv.1412.6980.

[42] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:190711692. 2019. https://doi.org/10.48550/arXiv.1907.11692.

[43] NTCIR. Test collections | data/tools. https://research.nii.ac.jp/ntcir/data/data-en.html. [Accessed 5 February 2024].

[44] Wada S. BLUE benchmark with transformers. https://github.com/sy-wada/blue_benchmark_with_transformers. [Accessed 4 May 2020].

[45] Wada S. seiya/oubiobert-base-uncased · hugging face. https://huggingface.co/seiya/oubiobert-base-uncased. [Accessed 14 May 2020].

[46] Wada S. Trials of pre-trained BERT models for the medical domain in Japanese. https://github.com/ou-medinfo/medbertjp. [Accessed 4 May 2020].