A Pre-Training Technique to Localize Medical BERT and to Enhance Biomedical BERT

Shoya Wada (wada@hp-info.med.osaka-u.ac.jp)
Osaka University Graduate School of Medicine https://orcid.org/0000-0001-7055-1009
Toshihiro Takeda
Osaka University Graduate School of Medicine
Shiro Manabe
Osaka University Graduate School of Medicine
Shozo Konishi
Osaka University Graduate School of Medicine
Jun Kamohara
Osaka University Faculty of Medicine Graduate School of Medicine: Osaka Daigaku Daigakuin Igakukei Kenkyuka Igakubu
Yasushi Matsumura
Osaka University Graduate School of Medicine

Research article

Keywords: BERT, Deep Learning, Neural Language Model, Natural Language Processing, Biomedical text mining

Posted Date: November 12th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-103477/v1

Abstract

Background: Pre-training large-scale neural language models on raw texts has been shown to make a significant contribution to a strategy for transfer learning in natural language processing (NLP). With the introduction of transformer-based language models, such as Bidirectional Encoder Representations from Transformers (BERT), the performance of information extraction from free text by NLP has significantly improved for both the general domain and the medical domain; however, it is difficult for languages in which there are few publicly available medical databases with a high quality and a large size to train medical BERT models that perform well.

Method: We introduce a method to train a BERT model using a small medical corpus both in English and in Japanese. Our proposed method consists of two interventions: simultaneous pre-training, which is intended to encourage masked language modeling and next-sentence prediction on the small medical corpus, and amplified vocabulary, which helps with suiting the small corpus when building the customized corpus by byte-pair encoding. Moreover, we used whole PubMed abstracts and developed a high-performance BERT model, Bidirectional Encoder Representations from Transformers for Biomedical Text Mining by Osaka University (ouBioBERT), in English via our method. We then evaluated the performance of our BERT models and publicly available baselines and compared them.

Results: We confirmed that our Japanese medical BERT outperforms conventional baselines and the other BERT models in terms of the medical-document classification task and that our English BERT pre-trained using both the general and medical domain corpora performs sufficiently for practical use in terms of the biomedical language understanding evaluation (BLUE) benchmark. Moreover, ouBioBERT shows that the total score of the BLUE benchmark is 1.1 points above that of BioBERT and 0.3 points above that of the ablation model trained without our proposed method.

Conclusions: Our proposed method makes it feasible to construct a practical medical BERT model in both Japanese and English, and it has a potential to produce higher performing models for biomedical shared tasks.

1 Background

Pre-training large-scale neural language models on raw texts has been shown to make a tremendous contribution to a strategy for transfer learning in natural language processing (NLP). With the introduction of transformer-based language models, such as Bidirectional Encoder Representations from Transformers (BERT), the performance of information extraction from free text by NLP has significantly improved in the general domain [1, 2]. Meanwhile, many studies, such as those on BioBERT, clinicalBERT, and SciBERT, showed that additional pre-training of BERT models on a large domain-specific text corpus, such as biomedical, clinical, or scientific text, results in satisfactory performance in their specific text-mining tasks [3–5] .

Although we have high expectations for the localization of medical BERT models, significant barriers exist to realize the localization. There are few publicly available medical databases written in each native language other than English with high-quality and a large size sufficient to train BERT models. For example, in Japanese, a subscription is required to perform a cross-search of Japanese medical journals, and most articles are published only in the PDF format, thereby making it difficult to obtain a large medical corpus.

We propose simultaneous pre-training in which we distinguish between two types of corpora and combine them to create pre-training instances via our method, and we have developed BERT models (see Fig. 1). We first developed a medical BERT model using a small medical corpus in Japanese, and we show the improvement that our method offers over the conventional models for a medical document classification task. Second, we applied it in English, and we show that the performance of the model is close to that of published models. Third, we demonstrate that our approach enables us to develop a pre-trained model that outperforms BioBERT.

In particular, we make the following contributions:

(1) We propose a method that enables users to train a medical BERT model using a small corpus. Subsequently, we show that the localization of medical BERT is feasible using our method.

(2) Applying our method, we developed a pre-trained model using PubMed abstracts and released it as Bidirectional Encoder Representations from Transformers for Biomedical Text Mining by Osaka University (ouBioBERT). We compare the performance of

ouBioBERT with the existing BERT models in terms of the biomedical language understanding evaluation (BLUE) benchmark [4], and we confirm that our model has a higher performance.

2 Materials And Methods

Our models essentially have the same structures as that of BERT-Base. We begin with an overview of BERT and describe available models used for medical text-mining tasks. Next, we illustrate our method and refer to our models. Finally, we explain the fine-tuning to evaluate our models.

2.1 BERT: Bidirectional encoder representations from transformers

BERT [2] is a contextualized word-representation model based on masked language modeling (MLM), and it is pre-trained using bidirectional transformers [1]. There are two steps in the BERT framework: pre-training and fine-tuning. During pre-training, the model is trained on unlabeled large corpora. For fine-tuning, the BERT model is first initialized with pre-trained weights, and all the weights are fine-tuned using labeled data from the downstream tasks. We applied a minimal architectural modification to the task-specific inputs and outputs into BERT and fine-tuned all the parameters in an end-to-end manner.

2.1.1 Pre-training

BERT pre-training is optimized for two unsupervised classification tasks (Fig. 2). The first is MLM. One training instance of MLM is a single modified sentence. Each token in the sentence has a 15% chance of being replaced by a special token [MASK]. The chosen token is replaced with [MASK] 80% of the time, 10% with another random token, and the remaining 10% with the same token. The MLM objective is a cross-entropy loss on predicting the masked tokens.

The second task is next-sentence prediction (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. Negative instances are created by pairing segments from different documents. Positive and negative instances are sampled with equal probability. The NSP objective is designed to improve the performance of downstream tasks, such as natural language inference (NLI) [6], which require reasoning regarding the relationships between pairs of sentences.

While creating the training instances, we set a duplicate factor, which contributes to data augmentation, while pre-training BERT. It refers to the duplicating times of the instances created from an input sentence, where these instances originate from the same sentence but have different [MASK] tokens.

2.1.2 Vocabulary

To manage the problem of out-of vocabulary words, BERT uses a vocabulary from subword units generated by WordPiece [7], which is based on byte-pair encoding (BPE) [8], for the unsupervised tokenization of the input text. The vocabulary is built such that it contains the most frequently used words or subword units. A main benefit of pre-training from scratch is to leverage a domain-specific custom vocabulary. For example, *appendicitis*, a common disease name, is divided into four pieces ([app, ##end, ##ici, ##tis]) by BERT [2] and three pieces ([append, ##icit, ##is]) by SciBERT [9]. Table 1 compares the vocabularies used by BERT variants. We refer to the original vocabulary released with BERT as BaseVocab, which is based on general domain corpora. In this study, custom vocabularies were used by SciBERT and our models in English. Meanwhile, vocabularies based on BaseVocab were used by the other BERT variants.

Table 1
Comparison of common medical terms in vocabularies used by BERT variants.

Medical Term	Category	BERT	SciBERT	ouBioBERT (Ours)
stroke	disease	✓	✓	✓
malalia	disease	✓	✓	✓
bleeding	symptom	✓	✓	✓
seizure	symptom	✓	✓	✓
pulmonary	organ	✓	✓	✓
stomach	organ	✓	✓	✓
surgery	procedure	✓	✓	✓
prescription	procedure	✓	✓	✓
cocaine	chemical	✓	✓	✓
glucose	chemical	✓	✓	✓
osteoporosis	disease		✓	✓
edema	symptom		✓	✓
pancreas	organ		✓	✓
laparotomy	procedure		✓	✓
dexamethasone	chemical		✓	✓
appendicitis	disease			✓
jaundice	symptom			✓
duodenum	organ			✓
polypectomy	procedure			✓
codeine	chemical			✓

Note: a "✓" symbol indicates that the corresponding vocabulary has the medical term; otherwise, the term will be broken up into smaller subwords.

2.1.3 Pre-trained BERT variants

The standard BERT model has been reported not to perform well in specialized domains, such as biomedical or scientific texts [3, 9]. To overcome this limitation, there are two possible strategies: either additional pre-training on domain-specific corpora from an existing pre-trained BERT model or pre-training from scratch on domain-specific corpora. A main benefit of the former is that the computational cost of pre-training is lower than the latter. The main advantage of the latter, as mentioned, is the availability of its custom vocabulary; however, the disadvantage is that the pre-trained neural language model may be less adaptable if the number of documents in a specific domain is small.

BERT-Base is pre-trained using English Wikipedia and BooksCorpus [2]. The vocabulary is BaseVocab, and its size is 30K. We evaluated the uncased versions of this model for the general domain.

BioBERT is the first BERT model released for the biomedical domain [3]. BioBERT is initialized from BERT-Base and trained using PubMed abstracts. We used BioBERT v1.1, whose vocabulary is based on BaseVocab, for the evaluation.

ClinicalBERT is a clinically oriented BERT model. [5]. It is initialized from BioBERT v1.0 and trained with additional steps using MIMIC-III clinical notes [10].

SciBERT leverages unsupervised pre-training on a large multi-domain corpus, which consists of 18% of papers from the computer science domain and 82% from the broad biomedical domain [9]. We evaluated the SciBERT-Base-Uncased that utilizes the original vocabulary called SciVocab.

BlueBERT is published with the BLUE benchmark [4]. In this study, we evaluated BlueBERT-Base (P) and BlueBERT-Base (P + M), which were initialized from BERT-Base and pre-trained using only PubMed abstracts and using the combination of PubMed abstracts for 5M steps and MIMIC-III clinical notes for 0.2M steps, respectively.

Tohoku-BERT is a Japanese BERT model used for the general domain released by Tohoku University [11]. It was pre-trained using Japanese Wikipedia, and its vocabulary was obtained by applying BPE to the corpus.

UTH-BERT is a clinical BERT model in Japan published by the University of Tokyo [12]. It was developed with a huge amount Japanese clinical narrative text, and its vocabulary was built with consideration of segment words for diseases or findings in as large a unit as possible.

2.2 Our proposed method: simultaneous pre-training and amplified vocabulary

If we train a BERT model only on a small medical corpus, we must be aware of the possibility that overfitting may degrade its performance. We hypothesized that it can be avoided if we simultaneously train a BERT model using both the general-domain and medical-domain knowledge. This would be achievable by increasing the frequency of pre-training for MLM using documents of the medical domain rather than the general domain and using the negative instances of NSP in which a sentence pair is constructed by pairing two random sentences, each from a different document. To increase the number of combinations of documents and to enhance medical-word representations in the vocabulary, we introduce the following two interventions.

Simultaneous pre-training is a technique used to efficiently create pre-training data from a set of corpora according to size and to pre-train a neural language model, as illustrated in Fig. 3. Given that we pre-train a medical BERT model, core corpora correspond to small medical corpora, and satellite corpora are large general-domain corpora, such as Wikipedia.

In the original implementation, we first divided the entire corpus into smaller text files that can be processed using the memory in practice. Subsequently, the combinations of NSP are determined within each split file, and the duplicate factor is set to define the number of times the sentences are used; however, there are two problems. The first is that the duplicate factor is applied to the entire corpora of both core corpora and satellite corpora, and thus the smaller corpora remain relatively small. Therefore, pre-training using core corpora is less frequent than pre-training using satellite corpora. The second is that the combinations of NSP are limited to the file that was initially split (see Fig. 3.A).

For our method, both core corpora and satellite corpora are first divided into smaller different documents with the same size and then combined to create pre-training instances. When we combined them, we ensured that the documents in the core and satellite corpora would be comparable in terms of their file sizes and that the patterns of the combination would be diverse. Using this technique, more instances from core corpora were used than those from satellite corpora, and they were homogeneously mixed (see Fig. 3.B). Consequently, this intervention achieved a higher increase in the frequency of pre-training for MLM using documents of core corpora through the process of pre-training than the original method. Furthermore, it generates an increased number of different combinations of documents compared to the original method.

As depicted in Fig. 3, core corpora and satellite corpora were combined so that their proportion was equal, and a sufficient number of pre-training instances were created to train a BERT model.

The **amplified vocabulary** is a custom vocabulary to suit a small corpus. If we build a vocabulary with BPE without adjusting the corpus sizes of core and satellite corpora, most words and subwords would be derived from the satellite corpora, which are larger than the core corpora. To solve this problem, we amplified the core corpora and made the corpus size the same as that of the satellite corpora. Subsequently, we constructed the uncased vocabulary via BPE using tokenizers [13].

2.3 Our pre-trained models

We produced the following BERT-Base models to demonstrate our method. The corpora we used for our models are listed in Table 2.

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

Abbr.	Corpus	Number of words	Size (GB)	Domain
jpW	Japanese Wikipedia	550M	2.6	(jp) General
jpCR	Digital clinical references	18M	0.1	(jp) Medical
W	English Wikipedia	2,200M	13	(en) General
BC	BooksCorpus	850M	5	(en) General
sP	Small PubMed abstracts	30M	0.2	(en) Biomedical
fP	Focused PubMed abstracts	280M	1.8	(en) Biomedical
oP	Other PubMed abstracts	2,800M	18	(en) Biomedical
Notes: Japanese corpora are tokenized using MeCab [14]. **jp: Japanese; en: English.**				

BERT (prop jpCR + jpW: AmpVocab) is a Japanese medical BERT model pre-trained using our method. We used a reference for clinicians, which is "Today's diagnosis and treatment: premium," and it consists of 15 digital resources for clinicians in Japanese published by IGAKU-SHOIN Ltd. as a source of medical knowledge (abbreviated as jpCR) and Japanese Wikipedia (jpW) as that of general-domain knowledge. For comparison, four pre-trained models were prepared. Two are publicly available models: **Tohoku-BERT** and **UTH-BERT**. The others are **BERT (jpW/ jpCR: jpWVocab)**, which was initialized with Tohoku-BERT and trained for additional steps using jpCR, and **BERT (jpCR: CRVocab)**, which was pre-trained only using jpCR from scratch.

Next, to assess the feasibility of adapting our method in English, we empirically produced a limited corpus of clinically relevant articles from PubMed abstracts. PubMed comprises a large number of citations for biomedical literature from MEDLINE, and therefore its articles are a mix of those for clinical medicine and those of the life sciences. To simulate a small medical corpus in English, we constructed a medical corpus, denoted as sP, extracted from the PubMed abstracts by using their medical subject headings (MeSH) IDs, which can be converted to the corresponding tree number. The heuristic rules used to decide which articles to extract are shown in Table A1.

BERT (prop sP + W + BC: AmpVocab) is a pre-trained medical BERT model in English to ensure that we can build a well-performing model using a small medical corpus via our method. We used our sP corpus as a small medical source and BooksCorpus (BC) and English Wikipedia (W) as general corpora. **BERT (sP: BaseVocab)** and **BERT (W + BC/ sP: BaseVocab)** were trained for comparison. The former was pre-trained solely using sP from scratch, and the latter was initialized from BERT-Base and trained using sP for domain-specific adaptation similar to BioBERT [3].

ouBioBERT (prop fP + oP: AmpVocab) is an enhanced biomedical BERT model pre-trained from scratch using entire PubMed abstracts in which medical articles, especially those related to human diseases, are amplified using our method. Our approach boosts the amount of training on the target domain within the entire corpus. We investigated whether the BERT model trained via our method using PubMed abstracts that were closely related to human diseases (focused PubMed abstracts: fP) as a core corpus and using other PubMed abstracts (oP) as a satellite corpus would achieve better performance for biomedical text-mining tasks than those of other BERT models. We created fP and oP from entire PubMed abstracts using their MeSH IDs (see Table A1).

To clarify the difference between our pre-trained models and the published models, we refer to the published models as shown in Table 3.

Table 3
List of the names for the published models discussed in this study.

Model	Name
English	
BERT-Base	BERT (W + BC: BaseVocab)
BioBERT	BioBERT (W + BC/ P: BaseVocab)
clinicalBERT	clinicalBERT (W + BC/ P/ M: BaseVocab)
SciBERT	SciBERT (Sci: SciVocab)
BlueBERT (P)	BlueBERT (W + BC/ P: BaseVocab)
BlueBERT (P + M)	BlueBERT (W + BC/ P/ M: BaseVocab)
Japanese	
Tohoku-BERT	BERT (jpW: jpWVocab)
UTH-BERT	UTH-BERT (EMR: EMRVocab)

Notes: W: English Wikipedia; BC: BooksCorpus; P: PubMed abstracts; Sci: scientific texts; M: MIMIC-III clinical notes; jpW: Japanese Wikipedia; EMR: electronic medical record of the University of Tokyo Hospital.

2.4 Task-specific fine-tuning BERT

Given an input token sequence, a pre-trained language model generates an array of vectors in the contextual representations. A task-specific prediction layer is then placed on top to produce the final output for the task-specific application. Given the task-specific training data, the task-specific model parameters can be trained and the BERT model parameters fine-tuned by gradient descent using backpropagation. Figure 4 shows a general architecture of fine-tuning BERT models for downstream tasks. Input instance is first subjected to task-specific pre-processing and to the addition of special instance markers ([CLS], [SEP], etc.). The transformed input is then tokenized using the vocabulary of the neural language model and input into the neural language model. The sequence of vectors in contextual representations taken from the language model is then processed by a Featurizer module and input into a Predict module to produce its final output of the given task.

Three evaluations were made. First, we studied the performance of the Japanese medical BERT variants and some baseline models other than neural language models on a medical-document-classification task to confirm that our method could be used in Japanese. Second, we showed the scores of the BLUE benchmark of BERT (prop sP + W + BC: AmpVocab) and publicly available pre-trained BERT models with a single random seed to demonstrate the effectiveness of our method in English. Finally, we executed the BLUE benchmark with five different random seeds and compared the average score of ouBioBERT (prop fP + oP: AmpVocab) with those of BioBERT (W + BC/ P: BaseVocab), BlueBERT (W + BC/ P: BaseVocab), and BlueBERT (W + BC/ P/ M: BaseVocab), respectively, to show the potential of our method.

3 Downstream Tasks

3.1 Multiclass document classification task in Japanese

Because there is no shared task for medical-domain documents in Japanese, we created a multiclass document classification task using the medical topics in the MSD Manual for the Professional [16] and named it DocClsJp. It is comprised of 2,475 articles, which belong to one of 22 disease categories.

We used the first 128 tokens of each document as an input sentence and defined its disease category as a correct label. We employed five-fold stratified cross-validation to evaluate the results using the micro-averaged F1-score. To compare the BERT models, we also evaluated the performance of conventional methods for DocClsJp.

3.2 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 [4]. These 10 corpora are pre-existing datasets that have been widely used by the Biomedical Natural Language Processing community as shared tasks (see Table 4). We used a macro-average of F1-scores and Pearson scores to make comparisons among pre-trained BERT models as a total score. Moreover, to evaluate the change of the total score by our method in detail, we calculated the scores of the clinical and biomedical domains as a clinical score and a biomedical score, respectively. That is, a clinical score is a macro-average of MedSTS, ShARe/CLEFE, i2b2 2020, and MedNLI, and a biomedical score is that of BIOSSES, BC5CDR-disease/chemical, DDI, ChemProt, and HoC [15, 17–24].

In this section, we briefly describe each of the individual tasks and datasets in the BLUE benchmark. For more information, refer to [4].

Table 4
BLUE tasks (Peng, et al., 2019).

Corpus	Type	Task	Metrics	Domain
MedSTS [23]	Sentence pairs	Sentence similarity	Pearson	Clinical
BIOSSES [19]	Sentence pairs	Sentence similarity	Pearson	Biomedical
BC5CDR-disease [17]	Mentions	Named-entity recognition	F1	Biomedical
BC5CDR-chemical [17]	Mentions	Named-entity recognition	F1	Biomedical
ShARe/CLEFE [18]	Mentions	Named-entity recognition	F1	Clinical
DDI [22]	Relations	Relation-extraction	micro F1	Biomedical
ChemProt [15]	Relations	Relation-extraction	micro F1	Biomedical
i2b2 2010 [20]	Relations	Relation-extraction	micro F1	Clinical
HoC [21]	Documents	Document classification	F1	Biomedical
MedNLI [24]	Pairs	Inference	accuracy	Clinical

3.2.1 Sentence similarity

The sentence-similarity task is used to predict similarity scores based on sentence pairs. It can be handled as a regression problem. Therefore, a special [SEP] token is inserted between the two sentences, and a special [CLS] token is appended to the beginning of the input. The BERT encoding of [CLS] is used in the calculation of the regression score. We evaluated similarity using Pearson correlation coefficients.

BIOSSES is a small dataset consisting of 100 pairs of sentences selected from the Text Analysis Conference Biomedical Summarization Track Training Dataset, which contains articles from the biomedical domain [19].

MedSTS is a dataset consisting of sentence pairs extracted from Mayo Clinic's clinical corpus and was used in the BioCreative/OHNLP Challenge 2018 Task 2 as ClinicalSTS [23].

3.2.2 Named-entity recognition

The named-entity recognition task aims to recognize mention spans given in a text. It is typically considered a sequential labeling task. The BERT encoding of a sequence of a given token is used to predict a label of each token and to recognize mentions of entities of interest. We evaluated the predictions using the strict version of the F1-score. For disjoint mentions, all spans must also be strictly correct.

BC5CDR-disease/chemical is a dataset derived from the BioCreative V Chemical-Disease Relation corpus, which was produced to evaluate relation-extraction of drug-disease associated interactions [17]. We trained named-entity recognition models for disease (BC5CDR-disease) and disease (BC5CDR-chemical) individually.

The **ShARe/CLEF** eHealth Task 1 Corpus is a collection of clinical notes from the MIMIC II database [18]. Annotations are assigned to the disorders written on the clinical notes.

3.2.3 Relation-extraction

The relation-extraction task predicts relations and their types between the two entities mentioned in the sentences. Following the practice in the BLUE benchmark [4], we regard this task as a sentence-classification task by anonymizing target named entities in the sentence using pre-defined tags, such as @GENE$ and @CHEMICAL$ [3]. By replacing mentions of entities with dummy tokens, we can avoid overfitting by memorizing the entity pairs.

The DDI corpus was developed for the DDI Extraction 2013 challenge and consists of 792 texts selected from the DrugBank database and 233 other Medline abstracts [22].

ChemProt consists of PubMed abstracts with chemical-protein interactions between chemical and protein entities and was used for the BioCreative VI chemical-protein interaction Track [15].

The **I2b2 2010** shared task was developed for the 2010 i2b2/VA challenge to determine concepts, assertions, and relations in clinical texts. Annotations were given for the relationship between the medical problem and either the treatment, the examination, or the other medical problem.

3.2.4 Document multilabel classification

The multilabel-classification task predicts multiple labels from the texts. **HoC** (the Hallmarks of Cancer corpus) was annotated with 10 hallmarks of cancer to help develop an automatic semantic classifier of scientific literature [21]. The annotation to texts from PubMed abstracts was made at the sentence level. We followed the common practice and evaluated the example-based F1-score at the document level [4, 25, 26].

3.2.5 Inference task

The inference task aims to predict whether the relationship between the premise and hypothesis sentences is a contradiction, an entailment, or neutral. **MedNLI** is an expert annotated dataset for NLI in the clinical domain and consists of sentence pairs sampled from MIMIC-III [10]. We evaluated the overall accuracy to evaluate the performance.

4 Experimental Setups

For both pre-training BERT and fine-tuning it for downstream tasks, we leveraged the mixed-precision training, called FP16 computation, which significantly accelerates the computation speed by performing operations in the half-precision format. We used two NVIDIA Quadro RTX 8000 (48 GB) GPUs for pre-training, whereas a single one was used for fine-tuning.

4.1 Pre-training BERT

We modified the implementation released by NVIDIA [27], which enabled us to leverage FP16 computation, gradient accumulation, and a layer-wise adaptive moments based (LAMB) optimizer [28], and we trained our models using the implementation. The configuration of the pre-training was almost the same as that of BERT-Base unless stated otherwise.

For BERT (prop jpCR + jpW: AmpVocab) and BERT (jpCR: CRVocab), the maximum sequence length was fixed at 128 tokens, and the global batch size (GBS) was set to 2,048. Additionally, a LAMB optimizer with the learning rate (LR) of 7e−4 was used. We trained the model for 125K steps. The size of the amplified vocabulary was 32K. BERT (jpW/ jpCR: jpWVocab) was initialized from BERT (jpW: jpWVocab) and trained until the loss of MLM and NSP on the training dataset stopped decreasing. Additionally, we used a LAMB optimizer with the LR of 1e−4.

We used the same settings for BERT (prop sP + W + BC: AmpVocab) and BERT (sP: BaseVocab) as that of BERT (prop jpCR + jpW: AmpVocab). BERT (W + BC/ sP: BaseVocab) was initialized from BERT (W + BC: BaseVocab) and trained for 25K steps with the same settings of the maximum sequence length and GBS as that of BERT (jpW/ jpCR: jpWVocab).

For ouBioBERT, we followed the NVIDIA implementation. First, we set the maximum sequence length of 128 tokens and trained the model for 7,038 steps using the GBS of 65,536 and a LAMB optimizer with the LR of 6e−3. Subsequently, we continued to train the

model allowing the sequence length up to 512 tokens for an additional 1,563 steps to learn positional embeddings using the GBS of 32,768 and a LAMB optimizer with the LR of 4e–3. The size of the amplified vocabulary was 32K.

4.2 Fine-tuning BERT for downstream tasks

We mostly followed the same architecture and optimization provided in transformers for fine-tuning [13]. In all the settings, we set the maximum sequence length to 128 tokens and employed the Adam optimizer [29] for fine-tuning using the batch size of 32 and the LR of 3e–5, 4e–5, or 5e–5, respectively. The number of training epochs was set for each task, as listed in Table 5. For each dataset and BERT variant, we picked the best LR and number of epochs on the development set, and then we reported the corresponding test results.

Table 5
Range of the number of training epochs for each task/dataset.

Dataset	Number of epochs
MedSTS	{7, 8, 9, 10}
BIOSSES	{40, 50}
Named-entity recognition	{20, 30}
Relation-extraction	{5, 6, 7, 8, 9, 10}
HoC	{5, 10, 15}
MedNLI	{5, 6, 7, 8, 9, 10, 15}
DocClsJp	{3, 4, 5, 6, 7, 8, 9, 10}

4.3 The performance of the baseline in DocClsJp

To evaluate the performance of the baseline, several conventional methods were applied.

One of the classical methods for text classification tasks is to use Support Vector Machines (SVM) to classify documents with features obtained from them [30]. The features are based on TF-IDF, numerical statistics that indicate how important a word is in a text by scoring the words in the document, considering the corpus to which the document belongs.

Deep neural networks for text classification tasks used before the introduction of transformer-based language models include Convolutional Neural Networks (CNN) and bidirectional Long Short-Term Memory (biLSTM) with self-attention [31, 32]. We first learned the word embeddings of jpCR, a Japanese medical corpus, using fastText [33]. Consequently, we converted a sequence of words from the documents by the embeddings and fed it into their neural networks. The structure of their networks was prepared based on each architecture of their original papers, respectively [31, 32].

For the three baseline methods, the maximum length of the input was set to 128 to match the input of our BERT models. The optimal hyperparameters were searched by Optuna, a hyperparameter optimization software designed using the define-by-run principle [34] .

5 Results

Table 6 compares the F1-score of the model pre-trained using our method and those of the others on DocClsJp. The performance of BERT variants is higher than baseline models of the other three. Our results show a higher performance of BERT (prop jpCR + jpW: AmpVocab) than those of the other pre-trained models either constructed using known techniques or publicly released. The ablation tests showed that simultaneous pre-training is more effective than existing methods and that its performance is enhanced by modifying the vocabulary with our method.

Table 6
Test results on DocClsJp.

Model		F1-score
TF-IDF + SVM		68.8 (1.3)
CNN		77.3 (2.8)
biLSTM with SA		78.9 (1.8)
BERT (jpW: jpWVocab)		82.3 (1.9)
UTH-BERT (EMR: EMRVocab)		82.7 (1.1)
BERT (jpCR: CRVocab)		84.4 (2.6)
BERT (jpW/ jpCR: jpWVocab)		84.6 (2.6)
BERT (prop jpCR + jpW: AmpVocab)		
SimPT	*AmpV*	
✓	✓	**87.2 (1.3)**
✓		85.6 (2.4)

Notes: **The numbers are mean (standard deviation) obtained using five-fold stratified cross-validation. TF-IDF + SVM: Support Vector Machines with TF-IDF; CNN: Convolutional Neural Networks for sentence classification** [31]; **biLSTM with SA: bidirectional Long-Short Term Memory with self-attention** [32]; **SimPT: simultaneous pre-training; AmpV: Amplified vocabulary.**

Table 7 summarizes the performance of BERT (prop sP + W + BC: AmpVocab) as well as those of publicly available BERT variants in terms of the BLUE score. BERT (prop sP + W + BC: AmpVocab) outperforms BERT (W + BC: BaseVocab) and BERT (sP: BaseVocab) and is as effective as BERT (W + BC/ sP: BaseVocab). Its high performance is close to that of domain-specific BERT models.

Table 7
BLUE scores of our BERT variants compared with those of the existing pre-trained models.

Model	Total	MedSTS	BIOSSES	BC5CDR-disease	BC5CDR-chemical	ShARe/CLEFE	DDI	ChemProt	i2b2 2010	HoC	MedNLI
BERT (W + BC: BaseVocab)	54.8	52.1	34.9	66.5	76.7	56.1	35.3	29.8	51.1	78.2	67.0
BioBERT (W + BC/ P: BaseVocab)	**82.9**	85.0	**90.9**	85.8	93.2	76.9	80.9	73.2	74.2	85.9	83.1
clinicalBERT (W + BC/ P/ M: BaseVocab)	81.2	82.7	88.0	84.6	92.5	78.0	76.9	67.6	74.3	86.1	81.4
SciBERT (Sci: SciVocab)	82.0	84.0	85.5	85.9	92.7	77.7	80.1	71.9	73.3	85.9	**83.2**
BlueBERT (W + BC/ P: BaseVocab)	82.9	**85.3**	88.5	**86.2**	**93.5**	77.7	**81.2**	**73.5**	74.2	**86.2**	82.7
BlueBERT (W + BC/ P/ M: BaseVocab)	81.8	84.4	85.2	84.6	92.2	**79.5**	79.3	68.8	**75.7**	85.2	82.8
BERT (sP: BaseVocab)	77.5	79.7	75.2	84.0	90.4	75.5	75.1	63.2	68.8	85.4	77.8
BERT (W + BC/ sP: BaseVocab)	81.4	83.2	90.7	86.0	92.2	77.8	76.8	68.2	73.2	85.1	81.0
BERT (prop sP + W + BC: AmpVocab)	81.4	83.2	89.7	85.7	91.8	79.1	78.4	67.5	73.1	85.3	80.1

Notes: **The best scores are in bold, and the second best are underlined.**

Table 8 compares each summarized score of the ouBioBERT (prop fP + oP: AmpVocab) results on the BLUE benchmark and with those of BioBERT (W + BC/ P: BaseVocab), BlueBERT (W + BC/ P: BaseVocab), and BlueBERT (W + BC/ P/ M: BaseVocab), respectively. Of the four models, ouBioBERT (prop fP + oP: AmpVocab) demonstrates the best score of the total score (1.0 point improvement, as shown in Table 8). We also conducted ablation tests. Consequently, we found that our configuration used for pre-

training ouBioBERT contributes to the greatest improvement in performance, particularly in biomedical scores, and that simultaneous pre-training is especially successful in improving clinical scores. The detailed results are shown in Table A2.

Table 8
Performance of ouBioBERT and its ablation tests on the BLUE task.

Model		Total score	Clinical score	Biomedical score
BioBERT (W + BC/ P: BaseVocab)		82.8 (0.1)	80.1 (0.3)	84.6 (0.4)
BlueBERT (W + BC/ P: BaseVocab)		82.9 (0.1)	79.8 (0.2)	85.0 (0.1)
BlueBERT (W + BC/ P/ M: BaseVocab)		81.6 (0.5)	**81.0** (0.3)	81.9 (0.9)
BERT (prop fP + oP: AmpVocab)				
SimPT	*AmpV*			
✓	✓	**83.9** (0.2)	80.5 (0.2)	**86.1** (0.2)
✓		83.9 (0.3)	80.6 (0.1)	86.0 (0.4)
		83.6 (0.1)	80.2 (0.3)	85.8 (0.2)

Notes: **The numbers are mean (standard deviation) on five different random seeds. The best scores are in bold, and the second best are underlined. SimPT: simultaneous pre-training; AmpV: Amplified vocabulary.**

6 Discussion

We confirmed that the model trained via our method even using a small medical corpus was robust for the BLUE benchmark, and we demonstrated that our method could construct both localized medical BERT and enhanced biomedical BERT.

We first applied our method to the medical BERT in Japanese and evaluated it for a single task. In our experiment, BERT (prop jpCR + jpW: AmpVocab) outperformed both the baseline models and the other BERT variants. Furthermore, in the ablation study, we observed that the performance was enhanced by the customized vocabulary via our method. Interestingly, UTH-BERT (EMR: EMRVocab), which was designed for the clinical domain in Japanese, was as accurate as BERT (jpW: jpWVocab), which was for the general domain. This is likely because DocClsJp was a classification task for medical references and the corpus used for pre-training BERT (prop jpCR + jpW: AmpVocab) consisted of clinical references and was therefore similar to the domain of the task. Similar results have been observed in English in a comparison between BioBERT (W + BC/ P: BaseVocab) constructed from PubMed and clinicalBERT (W + BC/ P/ M: BaseVocab) using MIMIC-III clinical notes [3, 5]. This suggests that we could localize medical BERT in Japanese for clinical references from a small medical corpus via our method.

Next, to simulate the experiment in Japanese, we created BERT (prop sP + W + BC: AmpVocab) by combining a small medical corpus and large general corpora in English. It performed sufficiently for practical use; however, BERT (sP: BaseVocab), which was pre-trained only using Small PubMed abstracts, performed worse than BERT (prop sP + W + BC: AmpVocab), and BERT (W + BC/ sP: BaseVocab), which was initialized from BERT (W + BC: BaseVocab) and pre-trained only using Small PubMed abstracts, was equivalent to BERT (prop sP + W + BC: AmpVocab). This supports the effectiveness of our method in using a small corpus in English, although the results were slightly different than those of the experiments in Japanese. The most significant difference lies between the model created by

our method and that created by domain adaptation with additional pre-training. This could be attributed to the effect of a custom vocabulary in the Japanese medical domain. Japanese sentences are described using a larger number of different characters than that of English. Moreover, medical terms are significantly different than general-domain words. Therefore, our amplified vocabulary could affect the performance of our BERT models in Japanese more strongly than in English. Notably, our method could create a medical BERT model that performed as satisfactory as or even better than the existing methods and that can be versatile. Therefore, it might be applicable in other languages as well. Furthermore, our method may be applied to professional domains other than the medical domain.

Finally, we demonstrated that a high-performance, pre-trained model could be trained using our method with ouBioBERT. The ablation test identified that the configuration we used in the pre-training of BERT models was the most significant to the improvement in their scores. Previous studies have reported that larger batch sizes and longer steps for pre-training are effective in improving performance [28, 35]; therefore, our model is likely to benefit from them. Furthermore, our simultaneous pre-training achieved an improvement in the BLUE benchmark scores, especially in clinical scores, although we used only PubMed abstracts rather than clinical notes. We designed this intervention with the intention of increasing both the frequency of pre-training for MLM on documents of target corpora and the combinations of documents for NSP. We contend that both contributed to the enhanced performance of ouBioBERT; however, in previous studies, it has been reported that NSP does not improve the performance of BERT models in the general domain, though this has not been reported in the medical domain yet [35, 36]. Further research is expected.

This study has several notable limitations. First, we checked the robustness of our models on multiple tasks in English; however, we evaluated BERT (prop jpCR + jpW: AmpVocab) for a single task in Japanese. This is because there are no text-mining shared tasks in Japanese for the medical domain, and it is difficult to directly solve this problem. Second, we do not determine the contribution of an amplified vocabulary to the performance in a language other than Japanese. To identify the contribution, we must conduct additional studies, such as a different construction of BERT's vocabularies or other experiments in the other languages; however, it is highly computationally expensive and significantly time-consuming for our environment to verify the contribution.

7 Conclusions

We introduced a pre-training technique that consists of simultaneous pre-training and an amplified vocabulary. We showed that a practical medical BERT model can be constructed via our method using a small medical corpus in English and that it can then be applied in Japanese. Additionally, using ouBioBERT, we confirmed that a pre-trained model that outperformed the pre-existing models can be produced using our method in the biomedical domain. Our study results could help overcome the challenges of biomedical text-mining tasks both in English and in other languages.

Abbreviations

BC
BooksCorpus; BC5CDR:BioCreative V Chemical-Disease Relation corpus; BERT:bidirectional encoder representations from transformers; biLSTM:bidirectional long short-term memory; BLUE:biomedical language understanding evaluation; BPE:byte-pair encoding; CNN:convolutional neural networks; CRVocab:jpCR-based vocabulary; EMR:enlectronic medical record of the University of Tokyo Hospital; EMRVocab:EMR-based vocabulary; fP:focused PubMed abstracts; GBS:global batch size; jpCR:digital clinical references; jpW:Japanese wikipedia; jpWVocab:jpW-based vocabulary; LAMB:layer-wise adaptive moments based; LR:learning rate; M:MIMIC-III clinical notes; MeSH:medical subject headings; MLM:masked language modeling; NLI:natural language inference; NLP:natural language processing; NSP:next-sentence prediction; oP:other PubMed abstracts; P:PubMed abstracts; PDF:portable document format; prop:proposed method; Sci:scientific texts; simPT:simultaneous pre-training; sP:small PubMed abstracts; SVM:support vector machines; W:English wikipedia

Declarations

Acknowledgments

Not applicable.

Authors' contributions

SW designed the project, developed the models and the codes, and was a major contributor in writing the manuscript. YM acquired the financial support, and supervised the project. JK investigated and analyzed the baseline. TT, SM, SK, and YM provided substantial contributions during manuscript writing and revision. All authors read and approved the final manuscript.

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).

Availability of data and material

We made ouBioBERT and the source code for fine-tuning freely available at https://github.com/sy-wada/blue_benchmark_with_transformers. We also published the pre-trained weights of Japanese medical BERT models in this study for academic purpose at https://github.com/ou-medinfo/medbertjp. The dataset of DocClsJp analyzed during the current study is not publicly available due to the restriction of the secondary distribution of copyrighted works, but it is available from the corresponding author on reasonable request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I: **Attention is all you need**. In: *Adv Neural Inf Process Syst: 2017*. 5998–6008.
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*. 4171–4186.
3. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J: **BioBERT: a pre-trained biomedical language representation model for biomedical text mining**. *Bioinformatics* 2020, **36**(4):1234–1240.
4. 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*. 58–65.
5. Alsentzer E, Murphy J, Boag W, Weng W-H, Jindi D, Naumann T, McDermott M: **Publicly Available Clinical BERT Embeddings**. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop: 2019*. 72–78.
6. 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*. 632–642.
7. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K: **Google's neural machine translation system: Bridging the gap between human and machine translation**. *arXiv preprint arXiv:160908144* 2016.
8. 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*. 1715–1725.
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*. 3606–3611.
10. Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG: **MIMIC-III, a freely accessible critical care database**. *Scientific data* 2016, **3**:160035.

11. Suzuki M: **cl-tohoku/bert-japanese: BERT models for Japanese text.** *https://github.com/cl-tohoku/bert-japanese* accessed: 13-Jan-2020.
12. Kawazoe Y, Shibata D, Shinohara E, Aramaki E, Ohe K: **A clinical specific BERT developed with huge size of Japanese clinical narrative**. *medRxiv* 2020:2020.2007.2007.20148585.
13. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M: **HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv 2019**. *arXiv preprint arXiv:191003771* 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*. 230-237.
15. 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*. 141–146.
16. Merck: **MSD Manual for the Professional**. *https://www.msdmanuals.com/ja-jp/professional* accessed: 03-Dec-2019.
17. Li J, Sun Y, Johnson RJ, Sciaky D, Wei C-H, Leaman R, Davis AP, Mattingly CJ, Wiegers TC, Lu Z: **BioCreative V CDR task corpus: a resource for chemical disease relation extraction**. *Database* 2016:baw068.
18. Suominen H, Salanterä S, Velupillai S, Chapman WW, Savova G, Elhadad N, Pradhan S, South BR, Mowery DL, Jones GJ: **Overview of the ShARe/CLEF eHealth evaluation lab 2013**. In: *International Conference of the Cross-Language Evaluation Forum for European Languages: 2013*. Springer: 212–231.
19. Soğancıoğlu G, Öztürk H, Özgür A: **BIOSSES: a semantic sentence similarity estimation system for the biomedical domain**. *Bioinformatics* 2017, **33**(14):i49–i58.
20. Uzuner Ö, South BR, Shen S, Duvall SL: **2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text**. *Journal of the American Medical Informatics Association* 2011, **18**(5):552–556.
21. Baker S, Silins I, Guo Y, Ali I, Högberg J, Stenius U, Korhonen A: **Automatic semantic classification of scientific literature according to the hallmarks of cancer**. *Bioinformatics* 2016, **32**(3):432–440.
22. Herrero-Zazo M, Segura-Bedmar I, Martínez P, Declerck T: **The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions**. *Journal of biomedical informatics* 2013, **46**(5):914–920.
23. Wang Y, Afzal N, Fu S, Wang L, Shen F, Rastegar-Mojarad M, Liu H: **MedSTS: a resource for clinical semantic textual similarity**. *Language Resources and Evaluation* 2020, **54**(1):57–72.
24. 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*. 1586–1596.
25. Zhang M-L, Zhou Z-H: **A Review on Multi-Label Learning Algorithms**. *IEEE Transactions on Knowledge and Data Engineering* 2014, **26**(8):1819–1837.
26. Du J, Chen Q, Peng Y, Xiang Y, Tao C, Lu Z: **ML-Net: multi-label classification of biomedical texts with deep neural networks**. *Journal of the American Medical Informatics Association* 2019, **26**(11):1279–1285.
27. NVIDIA: **BERT For PyTorch**. *https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/LanguageModeling/BERT* accessed: 24-Jan-2020.
28. You Y, Li J, Reddi S, Hseu J, Kumar S, Bhojanapalli S, Song X, Demmel J, Keutzer K, Hsieh C-J: **Large batch optimization for deep learning: Training bert in 76 minutes**. In: *International Conference on Learning Representations: 2020*.
29. Kingma DP, Ba J: **Adam: A Method for Stochastic Optimization**. *arXiv preprint arXiv:14126980* 2014.
30. Joachims T: **Text categorization with Support Vector Machines: Learning with many relevant features**. In*: 1998; Berlin, Heidelberg*. Springer Berlin Heidelberg: 137-142.
31. Kim Y: **Convolutional Neural Networks for Sentence Classification**. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP): oct 2014; Doha, Qatar*. Association for Computational Linguistics: 1746-1751.
32. Lin Z, Feng M, Santos CNd, Yu M, Xiang B, Zhou B, Bengio Y: **A structured self-attentive sentence embedding**. *arXiv preprint arXiv:170303130* 2017.
33. Bojanowski P, Grave E, Joulin A, Mikolov T: **Enriching word vectors with subword information**. *Transactions of the Association for Computational Linguistics* 2017, **5**:135-146.
34. Akiba T, Sano S, Yanase T, Ohta T, Koyama M: **Optuna: A Next-generation Hyperparameter Optimization Framework**. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; Anchorage, AK, USA*.

Association for Computing Machinery 2019: 2623–2631.

35. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V: **Roberta: A robustly optimized bert pretraining approach**. *arXiv preprint arXiv:190711692* 2019.
36. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R: **Albert: A lite bert for self-supervised learning of language representations**. *arXiv preprint arXiv:190911942* 2019.

Figures

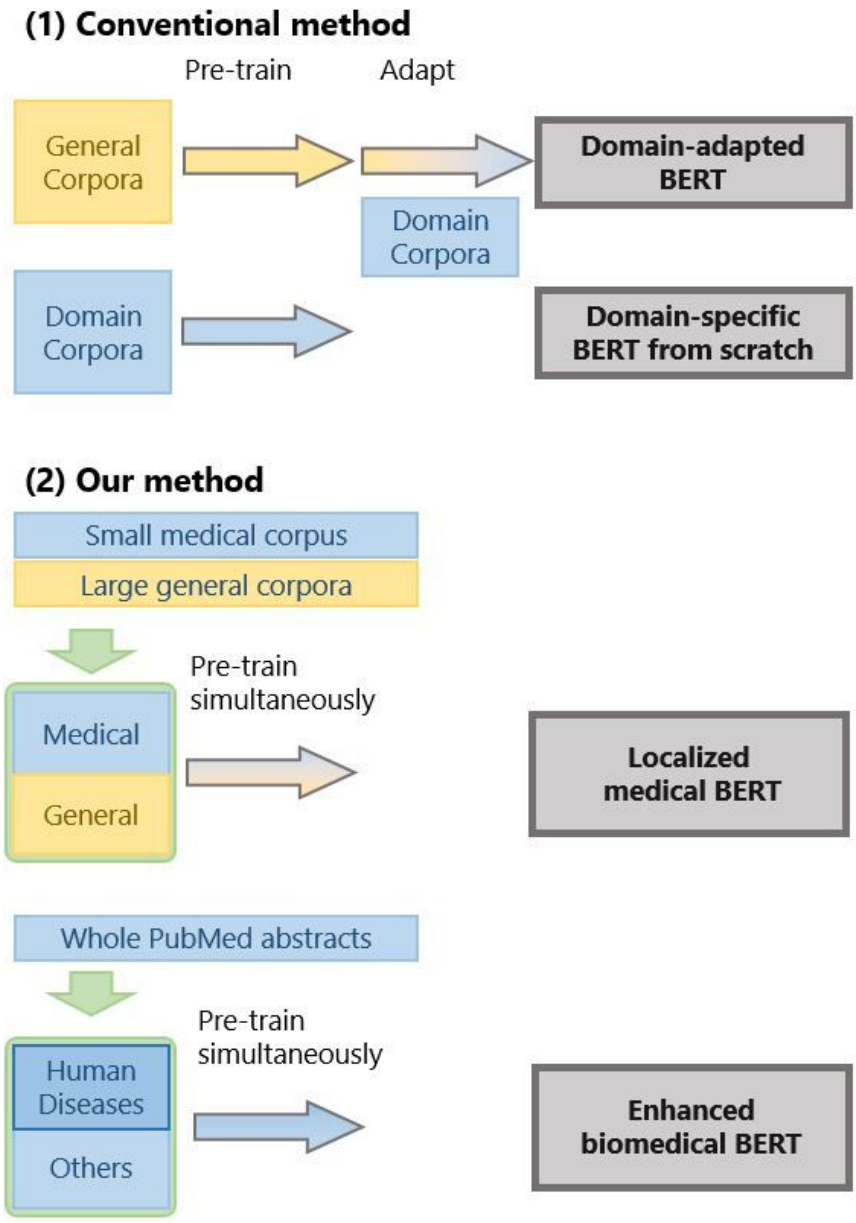

Figure 1

Overview of pre-training BERT.

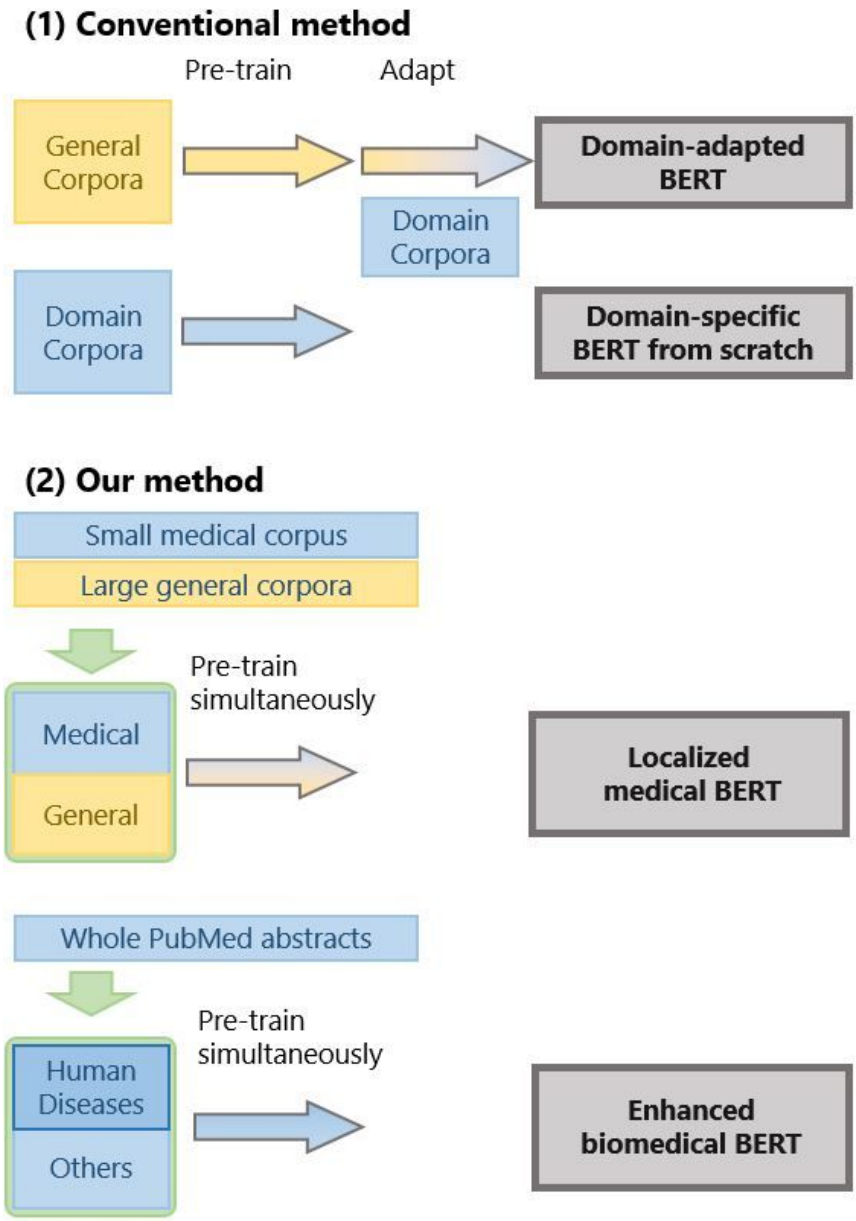

Figure 1

Overview of pre-training BERT.

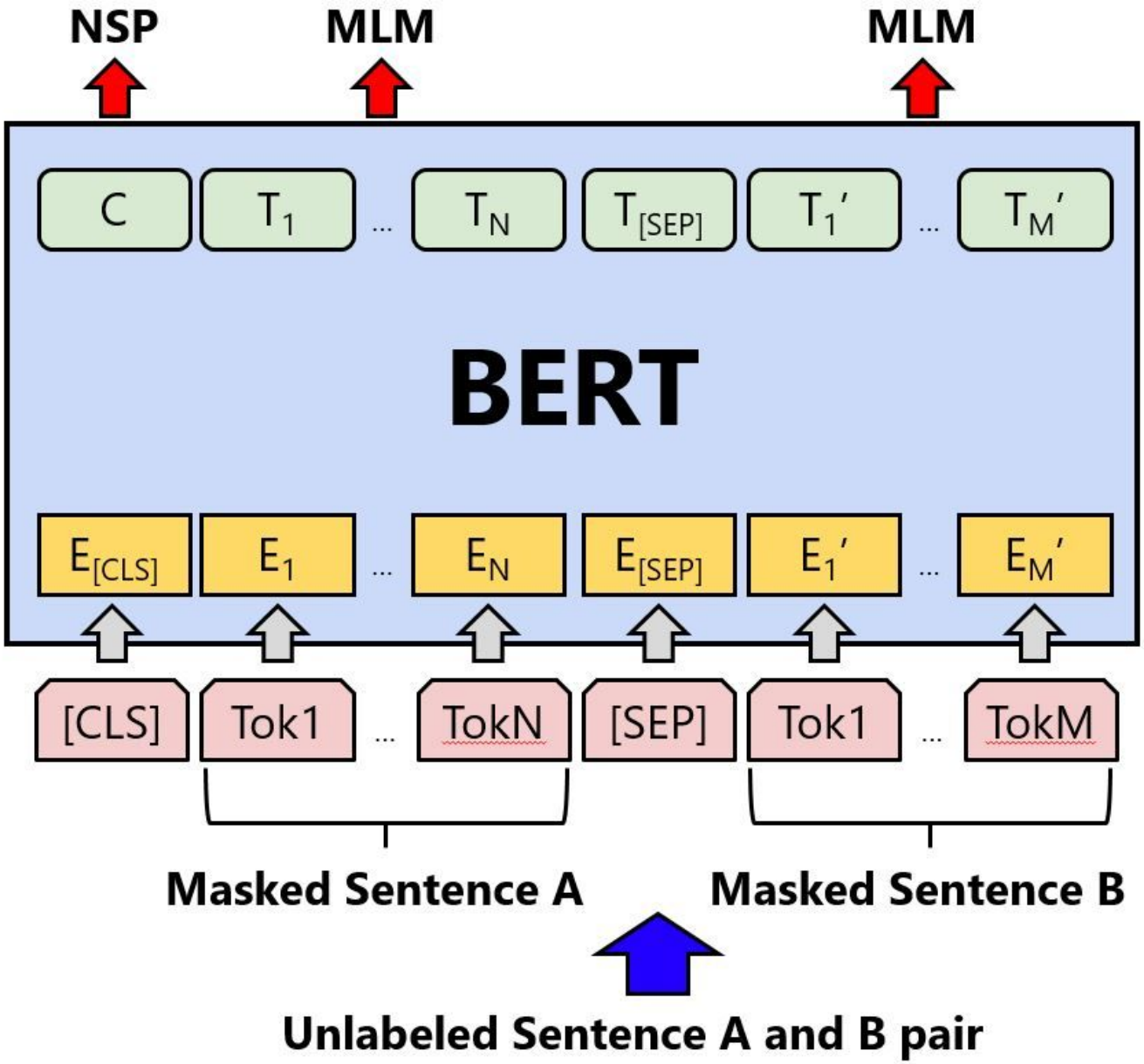

Figure 2

Pre-training procedures for BERT (adapted from Devlin 2019[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).

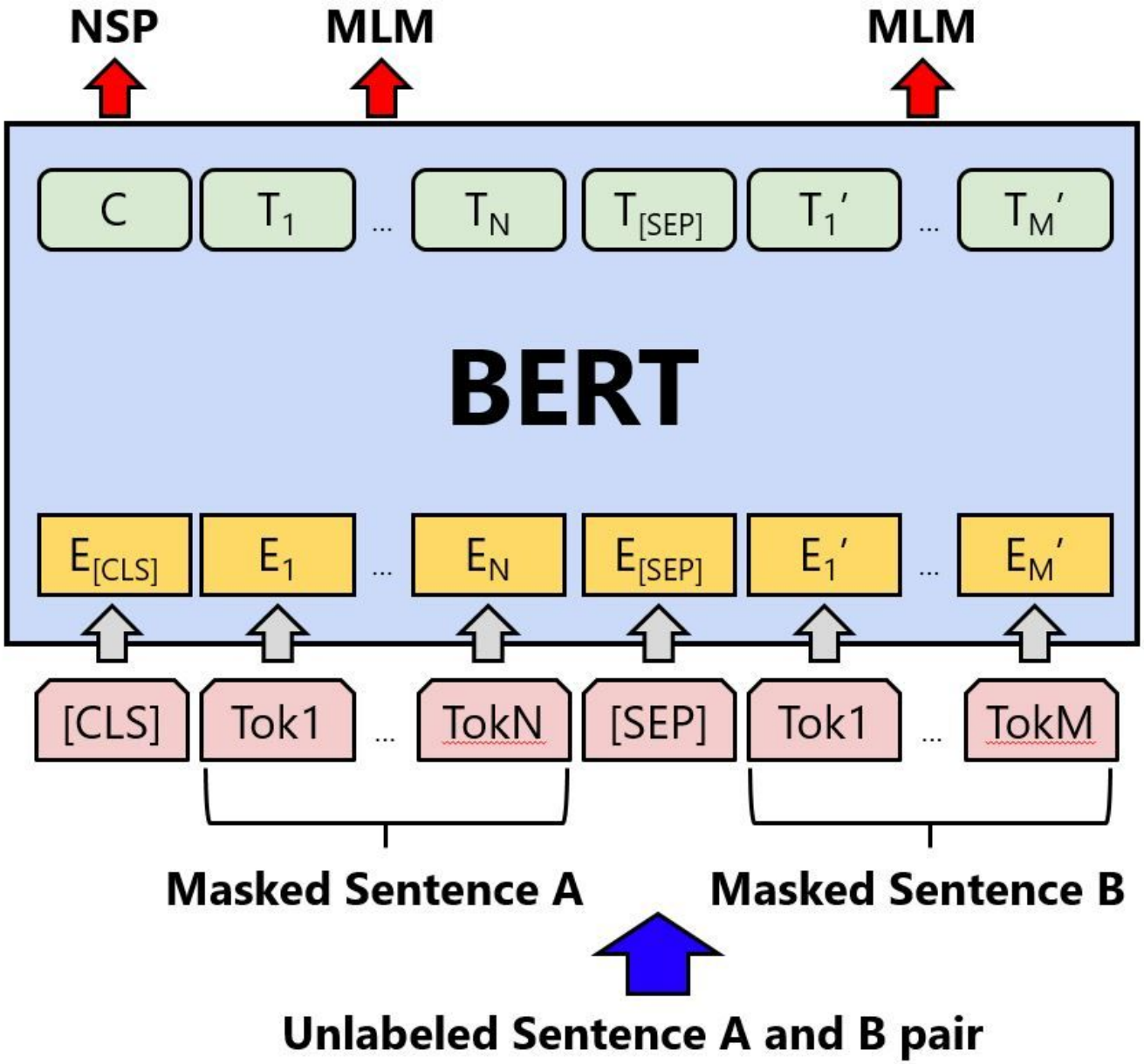

Figure 2

Pre-training procedures for BERT (adapted from Devlin 2019[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).

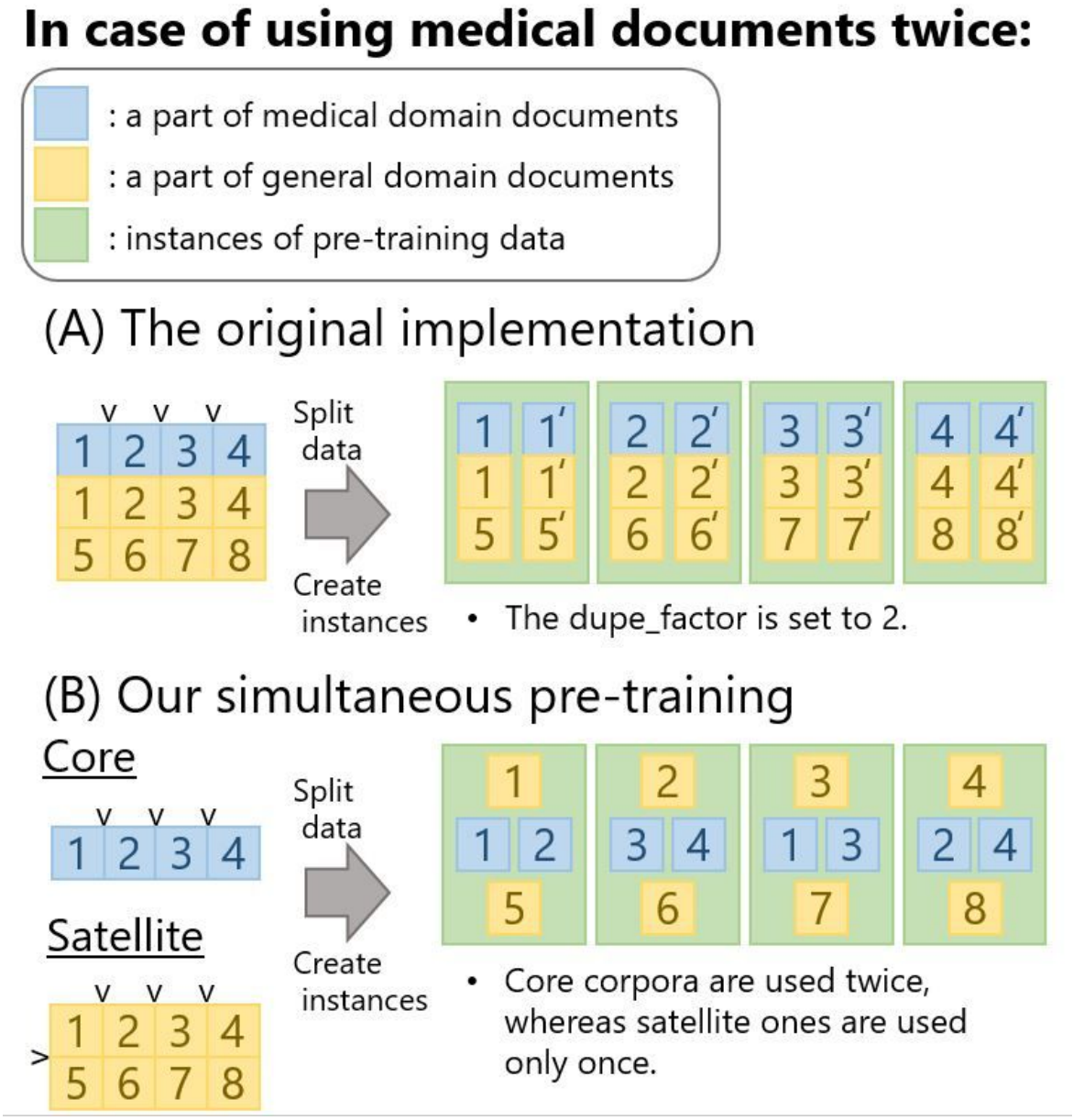

Figure 3

Our simultaneous pre-training method. (A) In the original implementation, documents are duplicated equally within each split group so that smaller corpora remain relatively small in the pre-training instances. (B) For our method, the amount of pre-training in-stances derived from smaller corpora, core, is apparently comparable with that from larger corpora, satellite.

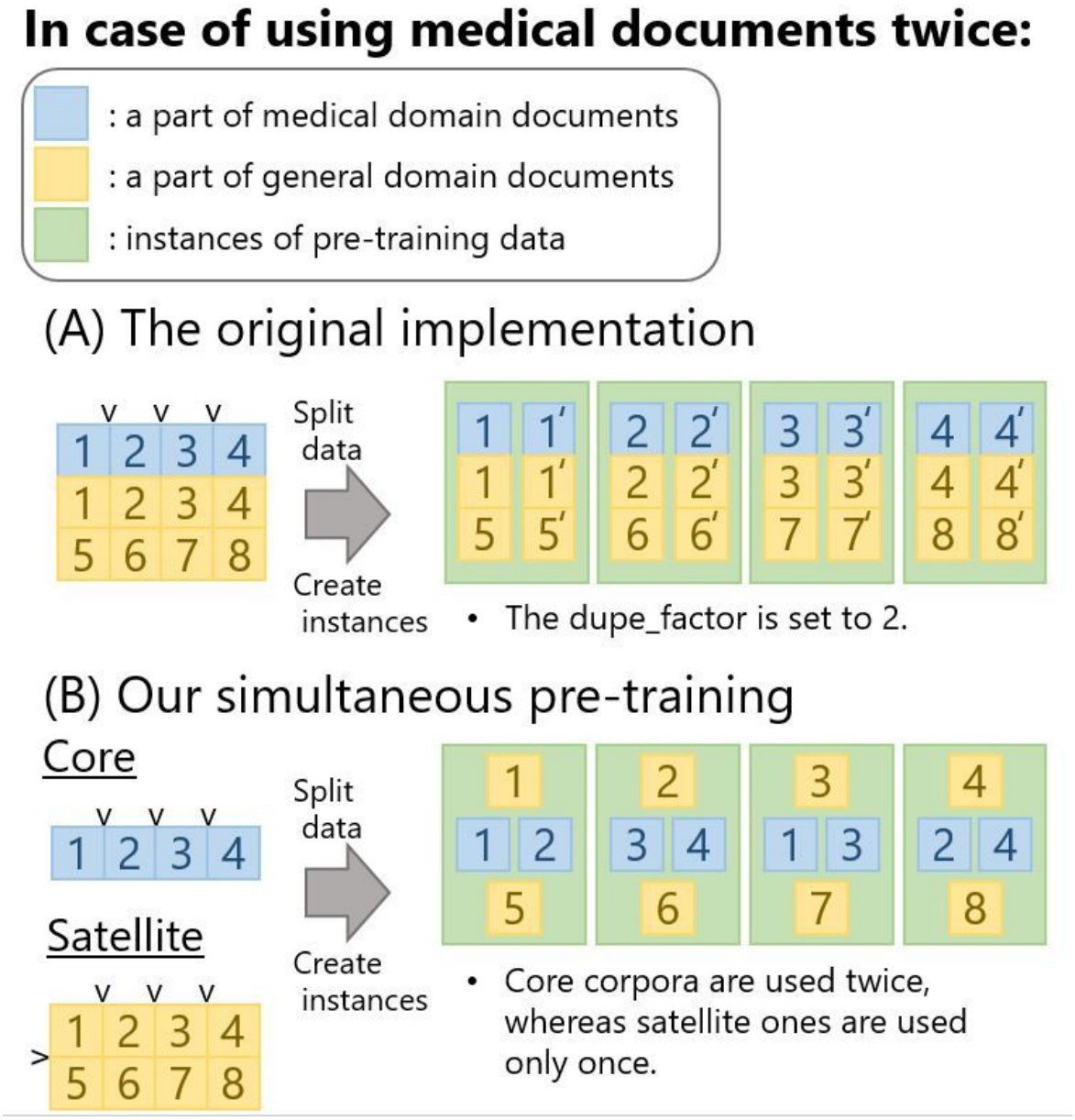

Figure 3

Our simultaneous pre-training method. (A) In the original implementation, documents are duplicated equally within each split group so that smaller corpora remain relatively small in the pre-training instances. (B) For our method, the amount of pre-training in-stances derived from smaller corpora, core, is apparently comparable with that from larger corpora, satellite.

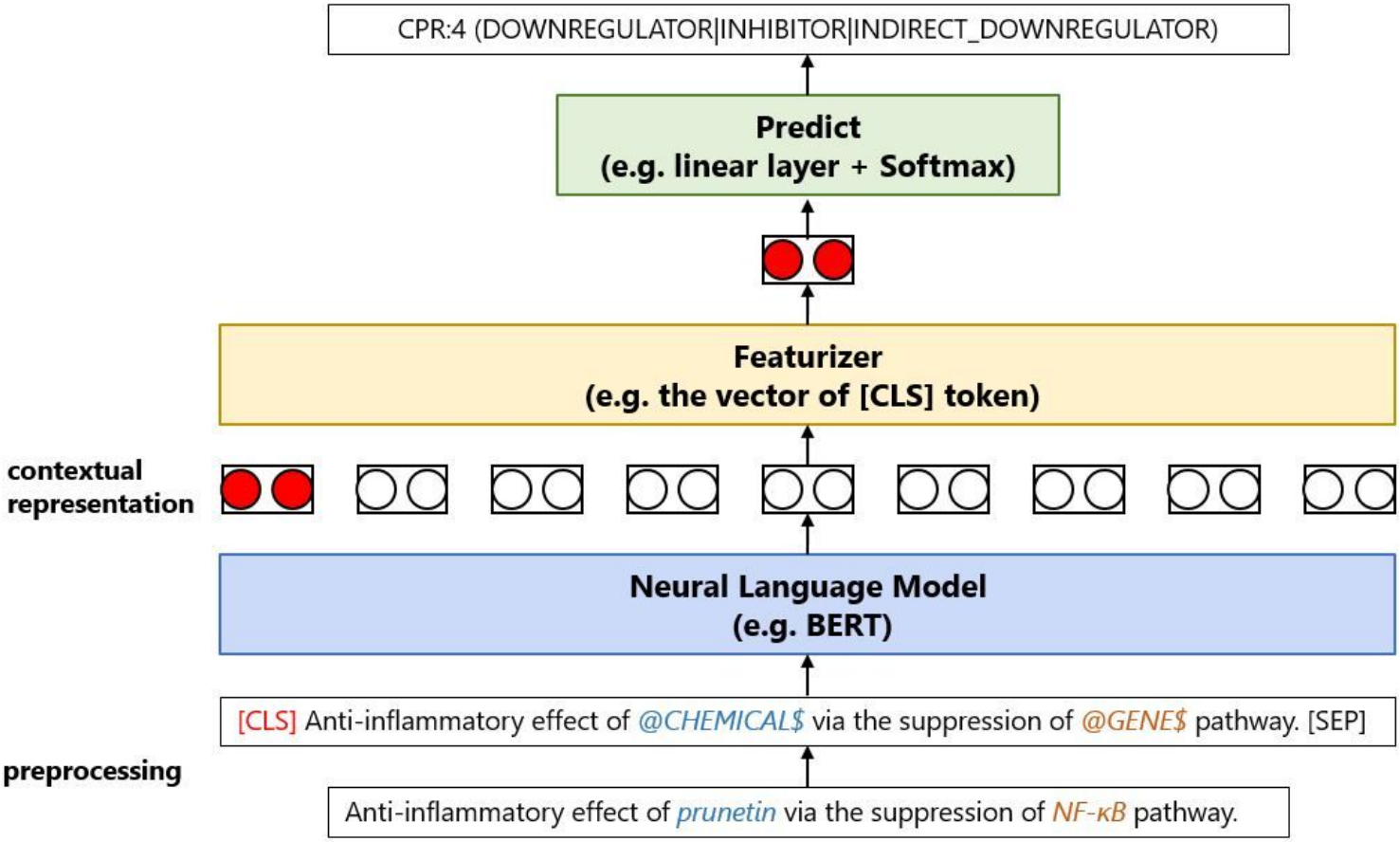

Figure 4

A general architecture of task-specific fine-tuning on neural-language models; an ex-ample of relation-extraction, ChemProt [15]. A sentence is transformed into an instance for BERT by replacing target entities with dummy tokens and adding special tokens. In a relation-extraction task, we use [CLS] BERT encoding as a Featurizer and predict the relationship between the entities by multiclass classification.

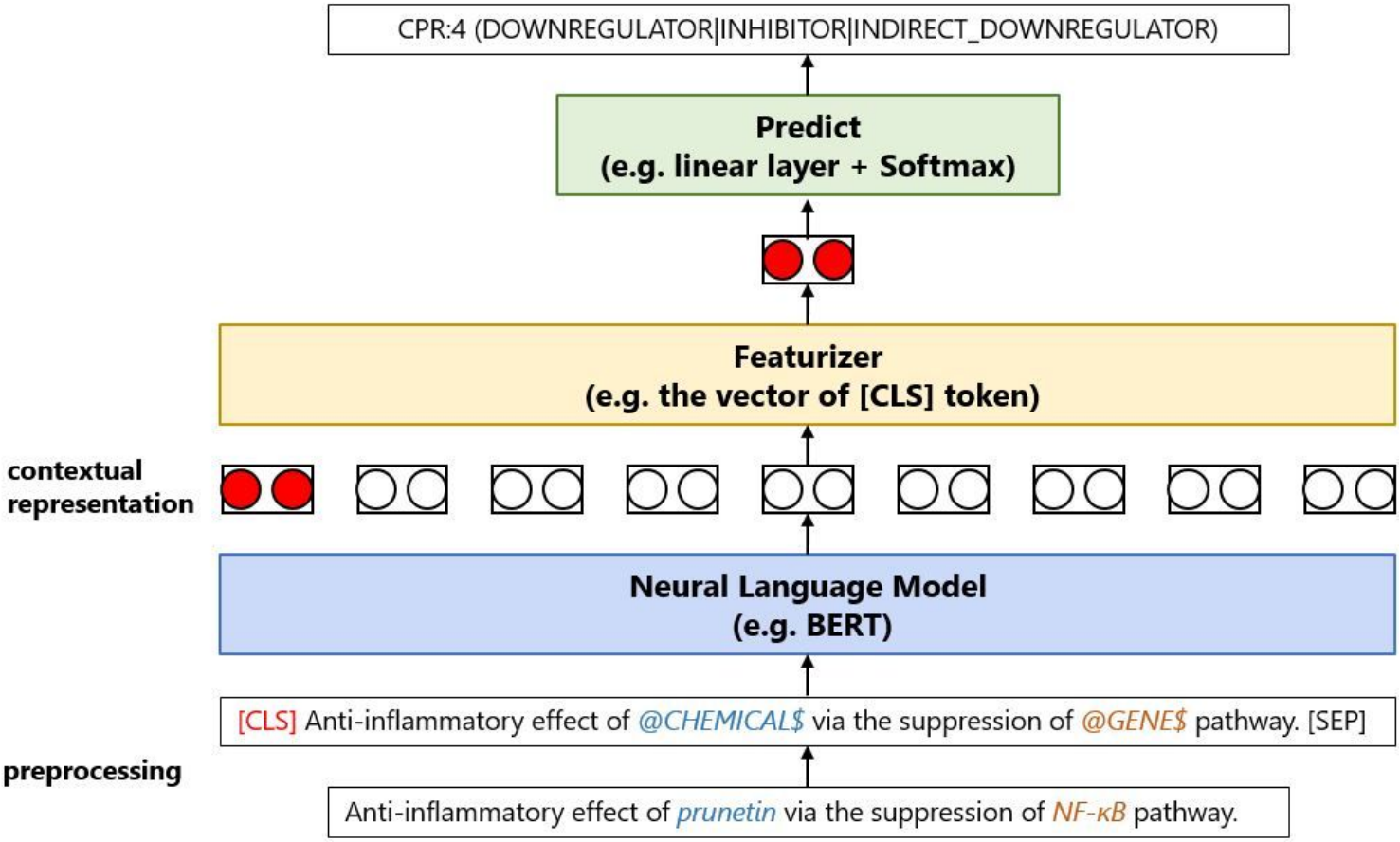

Figure 4

A general architecture of task-specific fine-tuning on neural-language models; an ex-ample of relation-extraction, ChemProt [15]. A sentence is transformed into an instance for BERT by replacing target entities with dummy tokens and adding special tokens. In a relation-extraction task, we use [CLS] BERT encoding as a Featurizer and predict the relationship between the entities by multiclass classification.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx
- Appendix.docx
- graphicalabstractfit.tif
- graphicalabstractfit.tif