Original Paper

Multifaceted Natural Language Processing Task–Based Evaluation of Bidirectional Encoder Representations From Transformers Models for Bilingual (Korean and English) Clinical Notes: Algorithm Development and Validation

Kyungmo Kim[1], MS; Seongkeun Park[2], MD, PhD; Jeongwon Min[1], MS; Sumin Park[3], PhD; Ju Yeon Kim[4], MD, PhD; Jinsu Eun[5], MS; Kyuha Jung[5], MS; Yoobin Elyson Park[5], MS; Esther Kim[5], BS; Eun Young Lee[4], MD, PhD; Joonhwan Lee[5], PhD; Jinwook Choi[3,6], MD, PhD

[1]Interdisciplinary Program for Bioengineering, Seoul National University, Seoul, Republic of Korea

[2]Seoul National University Medical Research Center, Seoul, Republic of Korea

[3]Institute of Medical and Biological Engineering, Medical Research Center, Seoul National University, Seoul, Republic of Korea

[4]Division of Rheumatology, Department of Internal Medicine, Seoul National University Hospital, Seoul, Republic of Korea

[5]Human Computer Interaction and Design Lab, Seoul National University, Seoul, Republic of Korea

[6]Seoul National University College of Medicine, Seoul, Republic of Korea

Corresponding Author:

Jinwook Choi, MD, PhD
Seoul National University College of Medicine
103 Daehak-ro, Jongno-gu
Seoul, 03080
Republic of Korea
Phone: 82 2-766-3421
Email: jinchoi@snu.ac.kr

Abstract

Background: The bidirectional encoder representations from transformers (BERT) model has attracted considerable attention in clinical applications, such as patient classification and disease prediction. However, current studies have typically progressed to application development without a thorough assessment of the model's comprehension of clinical context. Furthermore, limited comparative studies have been conducted on BERT models using medical documents from non–English-speaking countries. Therefore, the applicability of BERT models trained on English clinical notes to non-English contexts is yet to be confirmed. To address these gaps in literature, this study focused on identifying the most effective BERT model for non-English clinical notes.

Objective: In this study, we evaluated the contextual understanding abilities of various BERT models applied to mixed Korean and English clinical notes. The objective of this study was to identify the BERT model that excels in understanding the context of such documents.

Methods: Using data from 164,460 patients in a South Korean tertiary hospital, we pretrained BERT-base, BERT for Biomedical Text Mining (BioBERT), Korean BERT (KoBERT), and Multilingual BERT (M-BERT) to improve their contextual comprehension capabilities and subsequently compared their performances in 7 fine-tuning tasks.

Results: The model performance varied based on the task and token usage. First, BERT-base and BioBERT excelled in tasks using classification ([CLS]) token embeddings, such as document classification. BioBERT achieved the highest F_1-score of 89.32. Both BERT-base and BioBERT demonstrated their effectiveness in document pattern recognition, even with limited Korean tokens in the dictionary. Second, M-BERT exhibited a superior performance in reading comprehension tasks, achieving an F_1-score of 93.77. Better results were obtained when fewer words were replaced with unknown ([UNK]) tokens. Third, M-BERT excelled in the knowledge inference task in which correct disease names were inferred from 63 candidate disease names in a document with disease names replaced with [MASK] tokens. M-BERT achieved the highest hit@10 score of 95.41.

Conclusions: This study highlighted the effectiveness of various BERT models in a multilingual clinical domain. The findings can be used as a reference in clinical and language-based applications.

JMIR Med Inform 2024;12:e52897; doi: 10.2196/52897

Keywords: natural language processing; NLP; natural language inference; reading comprehension; large language models; transformer

Introduction

Since 2015, deep learning is increasingly being used in clinical natural language processing (NLP) [1]. Large language models (LLMs) based on deep learning technology are widely used in numerous clinical NLP domains [2]. Because contextual comprehension is critical for the overall performances of NLP models, studies have focused on the development of models that excel in conveying contextual information. Conventional approaches of NLP involve crafting word-to-word sequence models such as the hidden Markov model and using limited datasets annotated with labels such as disease and medication names [3-5]. However, studies are increasingly focusing on fine-tuning LLMs that have been pretrained on massive unlabeled biomedical literature sources, such as Medical Information Mart for Intensive Care (MIMIC-III) [6] and PubMed [7,8]. This shift in the NLP research direction has substantially elevated the contextual understanding capabilities of models and inspired studies on clinical NLP that focus on LLM utilization. For example, studies on automated summarization [9-11] have effectively extracted critical phrases from diverse sources, including biomedical papers and patient records. In addition, studies on entity extraction [12-15] have identified major entities such as disease names and drug names. However, these studies have focused exclusively on English-language corpora.

In the multilingual clinical domain, we proposed a set of contextual understanding conditions, with a comprehensive suite of clinical NLP evaluations specifically for these conditions. The proposed approach involves comparatively assessing bidirectional encoder representations from transformers (BERT) models [16] to provide guidelines for selecting the most suitable BERT model for a particular condition.

We proposed 2 hypotheses to examine 4 BERT models. First, we assumed that within the multilingual clinical domain, a language model capable of comprehending multiple languages would achieve superior performance. Second, models with the capacity to comprehend medical contexts would demonstrate superior efficacies. We selected BERT-base [16], Korean BERT (KoBERT) [17], Multilingual BERT (M-BERT) [18], and BERT for Biomedical Text Mining (BioBERT) [7] for the study. We pretrained these models on visit records on 160,000 patients. Subsequently, we introduced a series of comprehensive downstream tasks to learn the conditions required for these models to achieve effective contextual understanding. We assumed that an effective language model thrives in contextual comprehension under the following conditions:

- The model can determine whether the provided documents pertain to the same patient (tasks 1 and 2).
- The model is proficient in identifying the department associated with a given document (task 3).
- The model can discern the descriptions within medical records for the conditions of different patients (tasks 4 and 5).
- The model can ascertain the connection among sentences (task 6).
- The model can competently deduce disease names based on existing knowledge (task 7).

The rationale of the proposed conditions is the widespread adoption of BERT models in the medical domain for various applications.

BERT has been applied in medical natural language inference research to assess the relationship between 2 sequences (premise and hypothesis) with entailment, contradiction, or neutrality labels. Percha et al [19] used a fine-tuned BERT to locate clinical notes relevant to query sentences. Romanov and Shivade [20] created the MedNLI clinical dataset for natural language inference. They used several models and methodologies, such as bag-of-words, InferSent, and enhanced sequential inference models, to confirm the efficacy and validity of their datasets. Boukkour et al [21] introduced an alternative approach to BERT tokenization, proposing a convolutional neural network–based character-based tokenizer as a replacement for WordPiece Tokenizer, which is used to pretrain BERT, to improve BERT performance on the MedNLI dataset. Kanakarajan et al [22] pretrained the ELECTRA model, which is named "efficiently learning an encoder that classifies token replacements accurately," [23] using abstracts from PubMed, and evaluated its performance on the MedNLI dataset.

BERT was applied to categorize clinical notes. Rasmy et al [24] introduced Med-BERT, which pretrained BERT using electronic health record data to classify diabetes and pancreatic cancer datasets. This model exceeded gated recurrent units by 2-4 in terms of area under the receiver operating characteristic score. Zhang and Jankowski [25] proposed average pooling transformer layers handling token-, sentence-, and document-level embeddings for classifying *International Classification of Diseases* codes. Their model outperformed the BERT-base model by 11 points.

For the reading comprehension task, BERT can be used to determine the answer span within a given text. Pampari et al [26] proposed the electronic medical record question answering (emrQA) dataset to determine the answer span to a question in a clinical context. Yue et al [27] compared the performances of BERT-base, BioBERT, and ClinicalBERT [8] on the emrQA dataset and additional test datasets to address the problems of the emrQA dataset. Rawat et al [28] used 30 logical forms to express questions in semi-structured texts and identified the correct responses in the emrQA dataset. They entered clinical notes and questions and used multitask training to simultaneously predict the logical structure of the question and the text span of the answer in a clinical note. Savery et al [29] introduced the MEDIQA-AnS dataset, which contains questions and corresponding answers regarding the health care concerns of patients. The correct answers to these questions, which contain valuable information about the patients, are used as summaries.

BERT can be used to extract information from clinical notes. Yang et al [15] used the 2010 i2b2 [30], 2012 i2b2 [31], and 2018 national NLP clinical challenges (n2c2) [32] datasets to compare the information extraction performances of BERT models, namely, BERT-base, ELECTRA, A Lite BERT (ALBERT) [33], and Robustly Optimized BERT Pretraining Approach (RoBERTa) [34]. The test results revealed that RoBERTa outperformed the other models. Richie et al [35] used Clinical BERT [8] to extract the social determinants of patient health, namely, employment, living tobacco, alcohol, drug use, and their attributes, from the n2c2 2022 Track 2 dataset [36]; for instance, texts such as “works” and “unemployed” were extracted for detailing employment information.

Although studies have extensively examined BERT versatility, they have focused only on English corpora. To address this limitation, we comprehensively analyzed the efficacies of BERT models in various tasks involving medical documents in both Korean and English.

The rest of the manuscript is organized as follows. The *Methods* section outlines the diverse tests used for BERT analysis and their application procedures. The *Results* section presents a summary of the outcomes of each test. The *Discussion* section outlines the distinctive characteristics of each BERT model and presents a thorough analysis for understanding the reasons for these characteristics. Finally, the *Conclusion* section summarizes the study and emphasizes its significance.

The aim of this study was to identify the BERT models that perform optimally in the bilingual (Korean and English) clinical domain. To achieve this objective, we designed 7 tasks, evaluated the performance of 4 BERT variants (BERT-base, BioBERT, KoBERT, and M-BERT) across these tasks, and assessed their relative significance.

Methods

Dataset

We obtained outpatient records from 8 departments, namely, endocrinology, respiratory, cardiovascular, gastroenterology, rheumatology, nephrology, allergy medicine, and infectious medicine departments, at Seoul National University Hospital in South Korea. We collected the records of 164,460 outpatients between 2010 and 2019. The dataset comprised 2,453,934 documents, with 412,499,140 tokens generated after tokenization using white space. The distribution of tokens and documents for various departments was as follows: endocrinology (tokens: 91,352,271; docs: 496,938), respiratory (tokens: 31,556,578; docs: 195,048), cardiovascular (tokens: 114,978,554; docs: 696,061), gastroenterology (tokens: 57,755,571; docs: 416,062), rheumatology (tokens: 24,857,675; docs: 204,600), nephrology (tokens: 70,865,514; docs: 322,629), allergy medicine (tokens: 17,024,481; docs: 92,041), and infectious medicine departments (tokens: 4,108,496; docs: 30,555). Table 1 provides statistical data for the corpus. Table 2 presents the clinical note of a patient experiencing rheumatoid arthritis.

Table 1. Statistical data of clinical notes in Seoul National University Hospital between 2010 and 2019.

Department	Tokens, n	Documents, n
Endocrinology	91,352,271	496,938
Respiratory	31,556,578	195,048
Cardiovascular	114,978,554	696,061
Gastroenterology	57,755,571	416,062
Rheumatology	24,857,675	204,600
Nephrology	70,865,514	322,629
Allergy medicine	17,024,481	92,041
Infectious medicine	4,108,496	30,555
Sum	412,499,140	2,453,934

Table 2. The example of a clinical note that was used for training bidirectional encoder representations from transformers models (for better understanding, an English translation has been added).

Section	Contents
History	• Korean: *3117.2.1 arthralgia r/o d/t letrozole 로 병원 방문*; English (translated): *3117.2.1 arthralgia, rule out (r/o) due to letrozole. Visited hospital* • Korean: *meloxicam 7.5 mg bid 복용한 hx 있다.*; English (translated): *Has a history of taking meloxicam 7.5 mg twice daily.* • Korean: *f/u loss 마지막 방문 때 RF*[a]*, ACCP*[b]*, ANA*[c] *등 처방했다.*; English (translated): *Prescribed RF, ACCP, ANA, etc, during the last visit.*

Section	Contents
	• Korean: *Arthralgia , neutropenia 가 있다. 손이 붓고 마디가 아프다. 약먹지만 붓기가 빠지지 않는 것 같다.*; English (translated): *Experiencing arthralgia and neutropenia. Hands are swollen and joints are painful. Although taking medication, the swelling does not seem to be going down.*
P/E & Lab[d]	• Korean: *PIP S –/+ T –/– wrist S –/+ , T –/+ Toe s –/– T –/– 2120 . 3 lab bone scan: normal CBC*[e]*, WNL*[f]*, and CRP*[g] *0.10*; English (translated): *PIP S –/+ T –/– wrist S –/+ , T –/+ Toe s –/– T –/– 2120 . 3 lab bone scan: normal CBC, WNL, and CRP 0.10*
Assessment	• Korean: *Arthralgia r/o d/t letrozole 장상피화생*; English (translated): *Arthralgia r/o d/t letrozole. Intestinal metaplasia*
Plan	• Korean: *RF, ACCP, ANA , x-ray Celebrex 50 mg tid -->Celebrex 100 mg tid*; English (translated): *RF, ACCP, ANA, x-ray Celebrex 50 mg tid -->Celebrex 100 mg tid*

[a]RF: rheumatoid factor.

[b]ACCP: anticitrullinated protein antibody.

[c]ANA: antinuclear antibody.

[d]P/E & Lab: physical examination and laboratory.

[e]CBC: complete blood count.

[f]WNL: within normal limits.

[g]CRP: C-reactive protein.

Ethical Considerations

We obtained approval to use the original data collection for research purposes from the institutional review board (IRB) at Seoul National University Hospital (IRB no. C-2108-008-1242). According to the institution's IRB policy, the data cannot be publicly disclosed due to patient privacy concerns. Instead, we provide an overview of the data in Table 2.

BERT Models

The BERT-base model is a precursor in pretrained transformer encoders [37]. Vast open-domain data sources, including Wikipedia and BooksCorpus, are used to train the model [38]. The model is primarily focused on English text. The configuration of this dataset facilitates the expression of contextual representations of English sequences.

The BioBERT model is an evolution of BERT and is pretrained on PubMed data and enriched with biomedical entities, rendering BioBERT proficient in comprehending terminologies such as disease and drug names. In this study, we used the latest iteration of BioBERT, that is, BioBERT version 1.1.

The SKT Corporation in South Korea devised the KoBERT model to enhance the comprehension and processing of the Korean language. Data from Korean Wikipedia and news articles were used to pretrain the model.

The M-BERT model was obtained from a richly varied corpus of 104 languages, enabling a contextual representation that spans both English and Korean sequences.

Pretraining

To enhance the bilingual clinical contextual understanding capabilities of BERT models, we conducted additional pretraining using an extensive dataset comprising 159,460 out of 164,460 outpatient records from Seoul National University Hospital, employing masked language modeling. The data were preprocessed meticulously using this strategy. WordPiece Tokenizer was used by BERT-base, BioBERT, and M-BERT; SentencePiece Tokenizer [39] was used by KoBERT. All tokenizers were case-sensitive. Subsequently, random tokens within the input sequence were replaced with [MASK] tokens. This process was reiterated 10 times to yield the data required for pretraining. The pretraining task of the model involved reinstating the [MASK] token to its original token, drawing on the data crafted through this preprocessing procedure.

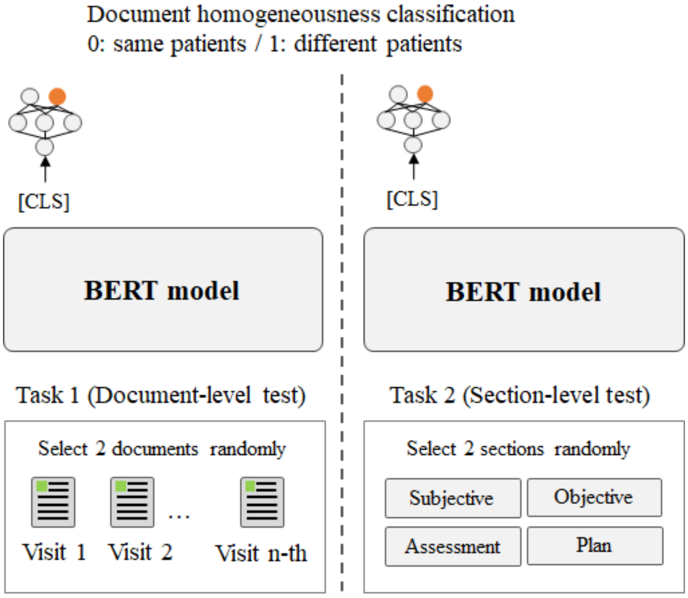
Multifaceted Clinical NLP Tasks

The evaluation framework encompassed 5 characteristics. Each characteristic was examined through 7 distinct downstream tasks that were designed to assess the clinical contextual comprehension capabilities of various BERT models.

Homogeneity Determination

As seen in Figure 1, we used 2 single outpatient records per input sequence to determine document homogeneity. Each model performed binary classification, discerning whether the records corresponded to those of the same patient (task 1). We extended this examination to the section level, tasking each model with predicting homogeneity based on a smaller segment of a page (task 2). In task 2, the objective was to determine whether 2 sequences originated from the same patient record, with 1 sequence containing an assessment section and the other section containing a randomized section.

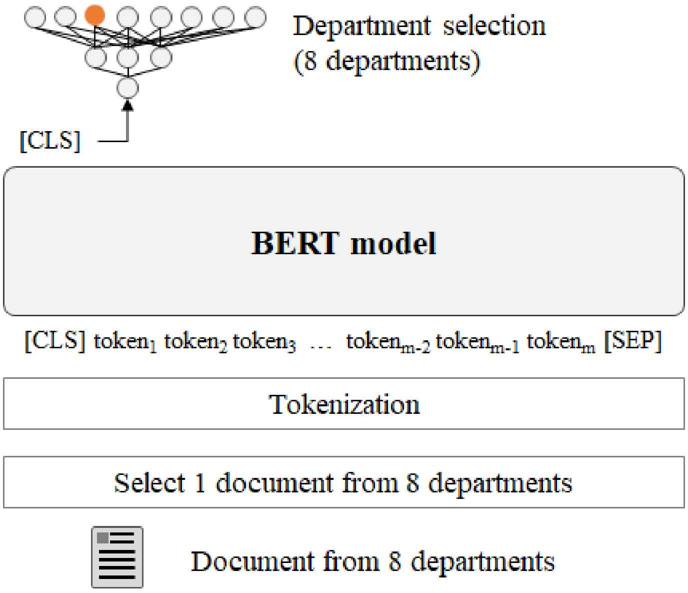
Figure 1. Document homogeneousness test (tasks 1 and 2). BERT: bidirectional encoder representations from transformers; CLS: classification.


Document homogeneousness classification
0: same patients / 1: different patients
[CLS]
[CLS]
BERT model
BERT model
Task 1 (Document-level test)
Task 2 (Section-level test)
Select 2 documents randomly
Visit 1
Visit 2
...
Visit n-th
Select 2 sections randomly
Subjective
Objective
Assessment
Plan

Document Representativeness

As seen in Figure 2, to assess document representativeness, we devised a task that focused on department identification by using individual visit records (task 3).

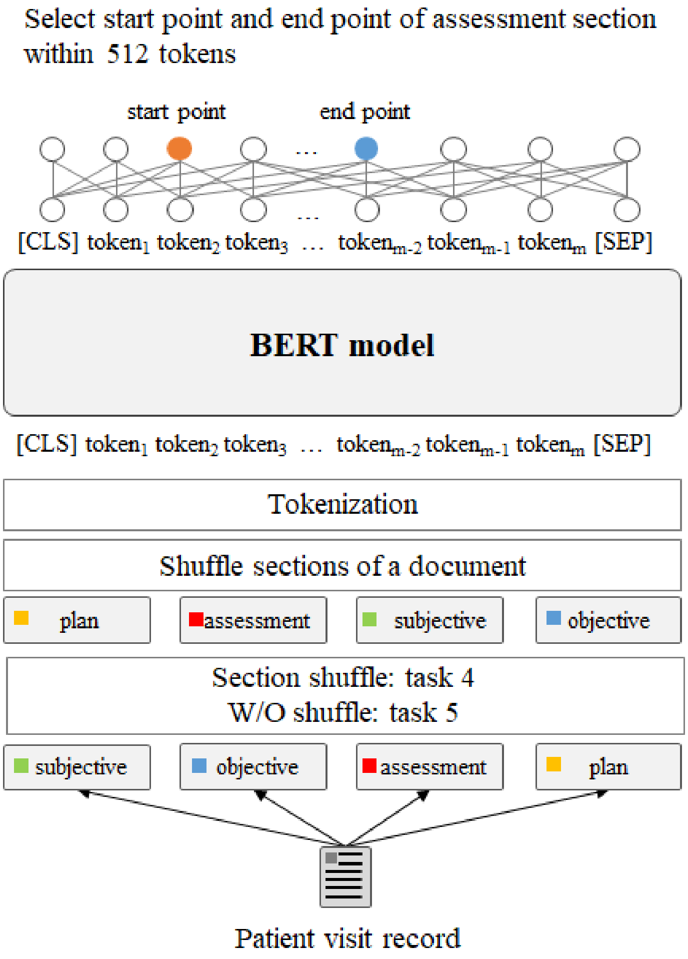
Figure 2. Document representativeness test: classifying documents (task 3). BERT: bidirectional encoder representations from transformers; CLS: classification; SEP: separator.


Department selection
(8 departments)
[CLS]
BERT model
[CLS] token1 token2 token3 ... tokenm-2 tokenm-1 tokenm [SEP]
Tokenization
Select 1 document from 8 departments
Document from 8 departments

Reading Comprehension Test

The reading comprehension test (Figure 3) test extracted summarized content from a visit record. We focused on extracting the assessment section from the Subjective, Objective, Assessment, Plan (SOAP) or the history, physical examination, laboratory, assessment, and plan sections. The experiments encompassed 2 setups, namely, 1 setup with section-shuffled documents (task 4) and 1 setup with maintained section-order documents (task 5).

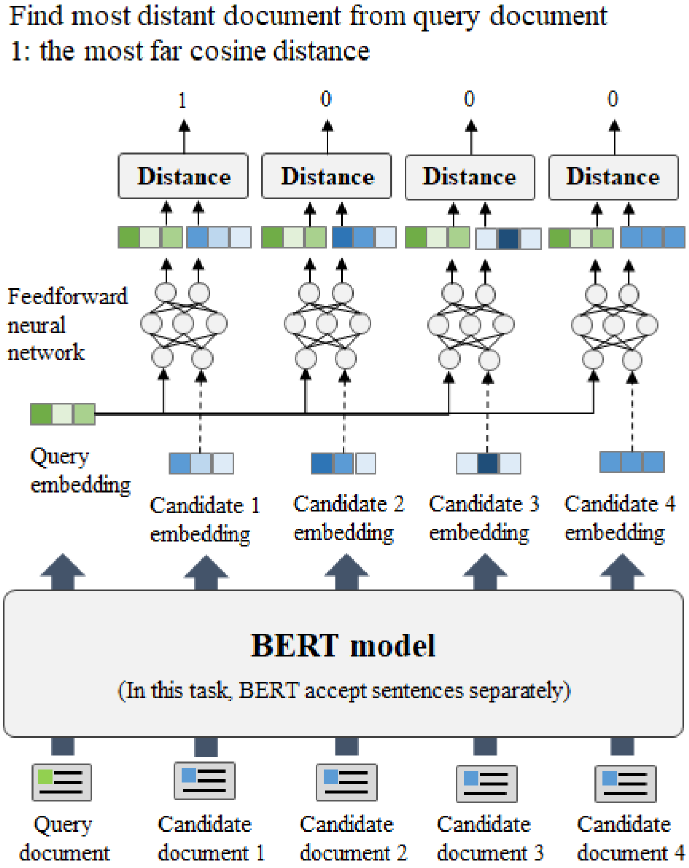
Figure 3. Reading comprehension test: identifying the department associated with a given document (with section shuffling: task 4; w/o section shuffling: task 5). BERT: bidirectional encoder representations from transformers; CLS: classification; SEP: separator; w/o: without.


Select start point and end point of assessment section within 512 tokens
start point
end point
[CLS] token1 token2 token3 … tokenm-2 tokenm-1 tokenm [SEP]
BERT model
[CLS] token1 token2 token3 … tokenm-2 tokenm-1 tokenm [SEP]
Tokenization
Shuffle sections of a document
plan
assessment
subjective
objective
Section shuffle: task 4
W/O shuffle: task 5
subjective
objective
assessment
plan
Patient visit record

Contextual Connections

As seen in Figure 4, we introduced a task that required the model to differentiate the most recent visit record from a set of 4 candidate documents when given a query document representing the oldest visit record (task 6). The limitation of BERT models regarding the amount of the input length they can handle necessitates a workaround because simultaneously inputting both the query document and 4 candidate documents is not feasible. To address this problem, we adopted a 2-step approach. First, each individual document was independently inputted into BERT to acquire document embeddings. Subsequently, these document embeddings, forming a pair comprising the query document and the kth candidate document embeddings, were introduced into a feedforward neural network (FFNN) [40]. For example, if the query and document embedding pair for the most recent visit were inputted into the FFNN, the model was trained to output a prediction value of 1; this value was assigned based on our assumptions. We postulated that the query document, which corresponded to the earliest visit among the 5 documents, and the last document, which denoted the most recent visit, encompassed the most distinct narrative. Consequently, we measured the cosine distances between these 2 embeddings and directed the model to output a prediction value of 1, which indicated the greatest distance in terms of cosine similarity. By contrast, if the query and nonanswer document embedding pairs were presented to the FFNN, the model was trained to output a prediction value of zero.

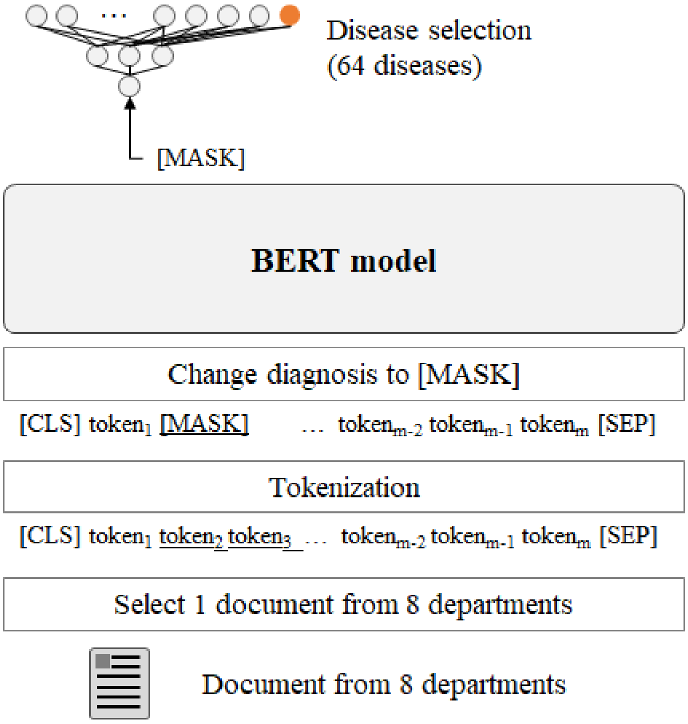
Figure 4. Document connectivity test: finding the last visited document (task 6). BERT: bidirectional encoder representations from transformers.



Knowledge Reasoning

The knowledge reasoning characteristic (Figure 5) evaluated the capacity of a model to deduce entities from masked text (task 7). Each model was tasked with deducing disease names from masked visit records in which the disease names had been replaced with [MASK] tokens. We used MetaMap [41] to create a dataset by identifying diagnostic names. Each model, when presented with the [MASK] token and context, selected the correct disease name from 63 disease names. A comprehensive list of the entities is shown in Table S1 in Multimedia Appendix 1.

Figure 5. Knowledge reasoning test: finding the disease name (task 7). BERT: bidirectional encoder representations from transformers; CLS: classification; SEP: separator.


Disease selection
(64 diseases)
[MASK]
BERT model
Change diagnosis to [MASK]
[CLS] token1 [MASK] … tokenm-2 tokenm-1 tokenm [SEP]
Tokenization
[CLS] token1 token2 token3 … tokenm-2 tokenm-1 tokenm [SEP]
Select 1 document from 8 departments
Document from 8 departments

Experimental Settings

We trained and evaluated 4 types of publicly available BERT models through the following process. We used records of 159,460 patients out of 164,460 patients for pretraining. In the pretraining procedure, 15% of random tokens from the 159,460 patient records were masked. Among them, 80% of the masked tokens were replaced with [MASK] tokens, 10% were replaced with random tokens, and the remaining 10% retained their original tokens. We trained the BERT models to restore [MASK] tokens to their original tokens.

After pretraining, the 4 BERT models were fine-tuned for tasks 1-7. For fine-tuning, we used 5000 patient records that were not used in pretraining. We assigned 4000 patients to the training set and 1000 patients to the test set and then created training and evaluation data specific for each task. In each task, the 4 pretrained BERT models were trained using the training set and evaluated on the test set.

In the pretraining step, 4 NVIDIA 3090 graphics processing units (GPUs) were used in parallel for 3 epochs. After pretraining, all the models were fine-tuned using a 1080ti GPU except for task 6, in which 3090 GPU were used, because this task required more calculation procedures and memory. The detailed hyperparameter settings are described in Table S2 in Multimedia Appendix 1. The detailed experimental settings and analysis code used in this study are available on GitHub [42].

Results

Results of Tasks 1-3

In tasks 1-3, BERT-base and BioBERT exhibited the best scores; Tables 3 and 4 present the corresponding results.

Table 3. Results of various BERT[a] models in tasks 1 and 2.

Model	Task 1: Determination of whether 2 documents are from the same patients			Task 2: Determination of whether 2 sections are from the same patients		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score
BERT-base	84.44	94.19	89.05	89.28	87.60	88.43
BioBERT[b]	83.36	96.21	89.32	92.92	82.73	87.53
KoBERT[c]	83.95	74.05	78.69	90.68	75.78	82.56
M-BERT[d]	83.22	94.02	88.29	83.56	93.38	88.19

[a] BERT: bidirectional encoder representations from transformers.
[b] BioBERT: BERT for Biomedical Text Mining.
[c] KoBERT: Korean BERT.
[d] M-BERT: Multilingual BERT

Table 4. Results of various BERT[a] models in task 3.

Model	Task 3: Identification of the department associated with a given document accuracy
BERT-base	96.75
BioBERT[b]	97.44
KoBERT[c]	95.38
M-BERT[d]	96.06

[a]BERT: bidirectional encoder representations from transformers.
[b]BioBERT: BERT for Biomedical Text Mining.
[c]KoBERT: Korean BERT.
[d]M-BERT: Multilingual BERT.

In the homogeneity test conducted on document-level inputs (task 1), BioBERT achieved the highest F_1-score, whereas in the test conducted on the section-level inputs (task 2), BERT-base achieved the highest F_1-score. Comparing the scores under tasks 1 and 2 revealed that BioBERT exhibited a more substantial drop in performance than those of other models. By contrast, KoBERT consistently demonstrated a diminished performance compared with that exhibited by other BERT models. In the document representativeness test, which entailed the selection of a single department from a set of 8 department candidates, BioBERT exhibited superior performance in terms of accuracy, which was the evaluation metric.

Results of Tasks 4-7

In tasks 4-7, M-BERT achieved the best scores (Tables 5 and 6).

Table 5. Results of various BERT[a] models in tasks 4 and 5.

Model	Task 4: Finding the assessment section with inputs that are section-shuffled			Task 5: Finding the assessment section with inputs that are not section-shuffled		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score
BERT-base	71.03	61.74	60.83	74.59	59.14	56.69
BioBERT[b]	72.16	56.31	51.64	74.71	55.99	51.17
KoBERT[c]	76.57	77.41	76.88	92.61	93.88	93.15
M-BERT[d]	93.15	94.61	93.77	96.52	96.37	96.44

[a]BERT: bidirectional encoder representations from transformers.
[b]BioBERT: BERT for Biomedical Text Mining.
[c]KoBERT: Korean BERT.
[d]M-BERT: Multilingual BERT.

Table 6. Results of various BERT[a] models in task 7.

Model	Task 7: Determination of disease names based on existing knowledge		
	hit@1	hit@3	hit@10
BERT-base[a]	60.26	77.47	93.40
BioBERT[b]	59.40	80.20	95.12
KoBERT[c]	46.20	72.02	91.54
M-BERT[d]	61.12	81.64	95.41

[a]BERT: bidirectional encoder representations from transformers.
[b]BioBERT: BERT for Biomedical Text Mining.
[c]KoBERT: Korean BERT.
[d]M-BERT: Multilingual BERT.

In the reading comprehension tests (tasks 4 and 5), the performances of the models were evaluated in terms of the F_1-score, which was calculated by measuring the proportion of tokens within the predicted interval that correctly overlapped with the actual interval. M-BERT achieved the highest performance in reading comprehension tests. In addition, the models exhibited the largest performance differences in these tests. In the context connectivity test (task 6), M-BERT exhibited the highest performance with an F_1-score of 64.75, whereas all the other models achieved a score lower than 60 (BERT-base: 59.78; BioBERT: 58.39; KoBERT: 25.62; and M-BERT: 64.75). In the knowledge-reasoning test (task 7), the M-BERT model exhibited the best performance. The primary objective of this test was to accurately prognosticate 63 potential candidate diagnoses, as extracted from clinical documents, in which the diagnosis name was substituted with [MASK]. In our assessment, we used hit@k (where k=1, 3, or 10). For instance, in task 7, BERT computes probabilities for 63 diseases based on a provided context. In this context, hit@k is a true positive if k diseases with the highest probability encompass the correct disease. The final evaluation score is then determined by dividing the number of true positives by the total number of sequences under assessment.

Discussion

Suitability of BERT-Base and BioBERT for English [CLS] Embedding (Tasks 1-3)

In tasks 1-3, the BERT classification ([CLS]) embedding was the input for the FFNN. The [CLS] token, positioned at the far-left side of the input sequence, is a classification token. The embedding of this token is commonly used as a feature for classification tasks, indicating the model's comprehension of segment-level or document-level context. In tasks 1 and 2, homogeneity was assessed at the document and section levels, respectively, and BioBERT and BERT-base demonstrated the highest performances, respectively. In task 3, BioBERT achieved the highest score. Based on these observations, we inferred that BERT-base and BioBERT would be suitable for tasks involving [CLS] embedding.

Generally, a model's ability to understand context diminishes as the number of tokens absent from its dictionary increases. Unknown ([UNK]) tokens represent tokens absent from the model's dictionary, and the presence of these tokens correlates with lower model performance. The higher the frequency of [UNK] tokens, the greater the challenge for the model to accurately comprehend the context. Notably, despite the limited inclusion of Korean tokens, these models excelled in tasks 1-3 (Table S4 in Multimedia Appendix 1). BERT-base and BioBERT, which were pretrained on English sentence patterns, exhibited improved performances because of the prevalence of English sentences in outpatient visit records, which typically detailed their diseases.

Influence of Multilingual Capabilities in Reading Comprehension Tasks on Outcomes (Tasks 4 and 5)

In tasks 4 and 5, the reading comprehension ability of the model was assessed by determining the scope of the assessment section. Among models, M-BERT demonstrated the highest performance, whereas BERT-base and BioBERT exhibited the lowest test scores. The presence of extensive multilingual capabilities in the reading comprehension tests was the predominant factor influencing these outcomes.

To comprehend why BERT-base and BioBERT exhibit markedly inferior performance compared with M-BERT in tasks 4 and 5, understanding the composition of the BERT model dictionaries and the function of the [UNK] token is crucial. In BERT models, a dedicated tokenizer is used to segment text into tokens. These tokens are retained if present in the model's dictionary; otherwise, the tokens are substituted with [UNK] tokens, representing unknown entities. Consequently, a higher prevalence of [UNK] tokens indicates a diminished ability of the model to comprehend the semantic nuances of the sequence. In tasks 4 and 5, where each token's semantic relevance determines its association with an assessment section, models with inadequate knowledge of individual tokens exhibit poor performance. The dictionaries of BERT-base and BioBERT contain minimal Korean characters, resulting in the majority of Korean tokens being replaced with [UNK] tokens. By contrast, M-BERT encompasses a comprehensive range of Korean characters in its dictionary. Therefore, BERT-base and BioBERT exhibit notably inferior performance in tasks 4 and 5 compared with M-BERT.

Relationship Between Multilingual Capability and Task Complexity (Task 7)

Task 7, which was focused at evaluating the aptitude of a model for knowledge inference, was more complex than other tasks. Notably, M-BERT outperformed the other models in task 7, securing hit@1, hit@3, and hit@10 scores of 61.12, 81.64, and 95.41, respectively. These results highlighted the pivotal role of the dictionary in knowledge inference. Furthermore, when processing documents in multiple languages, M-BERT outperformed BERT-base, which had been exclusively trained in a single language.

For task 6, the test results were poor. An analysis indicated that BERT models did not excel in this task because of the prevalence of outpatient medical records in the copy-and-paste format (Table S6 in Multimedia Appendix 1). Consequently, the significance of task 6 in this study was low.

Contributions to the Clinical Text Processing and Medical Fields

Importance of Multilingual Models

The experiment highlights the significance of using multilingual language models in processing bilingual clinical notes. The findings demonstrated that using a model capable of handling 2 languages yields superior performance compared with relying solely on a single language model. This insight is particularly relevant for countries such as Korea and Japan, where clinical documentation typically involves a mixture of languages.

Base for Model Selection

Furthermore, this study provides empirical evidence for choosing a proper BERT model, a factor not substantiated in existing NLP research. For instance, in previous studies, such as that conducted by Kim and Lee [43], M-BERT was used for tasks such as extracting disease names, symptoms, and body parts from Korean text without providing explicit justification. The experimental results satisfied this gap by showcasing the superiority of M-BERT in understanding bilingual clinical text and supporting appropriate BERT selection in future studies.

Limitations and Future Works

Limited Scope of Clinical Notes

This analysis primarily focused on outpatient visit records. Future studies should encompass a broad range of clinical notes, including surgical notes, hospitalization records, and discharge summaries. Comparing and validating the performance of BERT models across various types of clinical documentation provides a comprehensive understanding of their effectiveness.

Single-Institution Data

This study exclusively used data from Seoul National University Hospital, which can limit the generalizability of the findings. Clinical notes can vary considerably in style and content across various health care institutions. Therefore, future studies should involve data from multiple hospitals to validate BERT model performance in various clinical settings.

More Tasks Should Be Verified

The BERT model requires further validation in bilingual clinical text. Oh et al [44] conducted a study to recognize protected health information in the publicly available i2b2 2014 dataset. However, we could not perform this task because manual labeled annotations are required to extract non-English entities in bilingual clinical notes. In future studies, various tasks using bilingual clinical notes should be proposed.

Conclusions

In this study, we comprehensively compared 4 BERT models, encompassing text in both English and Korean, within the multilingual clinical domain. We pretrained these models with approximately 160,000 patient records and evaluated their performances for 7 diverse downstream tasks. The experimental findings are summarized as follows.

First, the BERT-base and BioBERT models excelled in document classification tasks using [CLS] tokens. These results highlighted their superiority over M-BERT in tasks involving simple pattern recognition in word sequences. Second, the significance of having a comprehensive dictionary was evident in the reading comprehension task in which comprehensive token usage was required. The exceptional performance of M-BERT, which encompassed a broad range of Korean and English tokens, clearly confirmed the importance of the dictionary. Third, multilingual proficiency was pivotal for tasks that demanded complex reasoning. Both M-BERT and BioBERT excelled in task 7, which focused on diagnosing a multitude of candidates, and notably, M-BERT consistently outperformed BioBERT.

Our findings highlighted the suitability of BioBERT and BERT-base for tasks that relied on sequence patterns in multilingual clinical domains. In addition, M-BERT, which had an expansive dictionary and aptitude for leveraging Korean and English clinical contexts, was highly suitable for tasks involving textual content comprehension. The experimental results of the BERT models in mixed-language clinical documents provide valuable insights for future medical NLP research and appropriate BERT model selection for different types of tasks.

Acknowledgments

This study was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (no. NRF-2021R1I1A4A01042182).

Data Availability

The datasets generated and/or analyzed during the current study are not publicly available due to patient privacy concerns. Patient records contain personal information, and, as such, Seoul National University's institutional review board does not permit public disclosure of the data.

Authors’ Contributions

KK, SP, JM, and SP conceptualized the paper, developed the methodology, and prepared the original draft of the manuscript. KK contributed to software implementation and validated the findings. JYK and EYL curated the data, conducted investigations, and contributed to data analysis and interpretation. JE, KJ, YEP, EK, and JL contributed to methodology development, conducted formal analysis, and provided insights throughout the research process. JC supervised the study, managed the project administration, and contributed to reviewing and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary materials on disease entities, hyperparameter settings, number of documents in the pretraining dataset, tokens, tokenization, and masked language modeling loss.
[DOCX File (Microsoft Word File), 125 KB-Multimedia Appendix 1]

References

1. Wu H, Wang M, Wu J, et al. A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. NPJ Digit Med. Dec 21, 2022;5(1):186. [doi: 10.1038/s41746-022-00730-6] [Medline: 36544046]
2. Karabacak M, Margetis K. Embracing large language models for medical applications: opportunities and challenges. Cureus. May 2023;15(5):e39305. [doi: 10.7759/cureus.39305] [Medline: 37378099]
3. Zhang J, Shen D, Zhou G, Su J, Tan CL. Enhancing HMM-based biomedical named entity recognition by studying special phenomena. J Biomed Inform. Dec 2004;37(6):411-422. [doi: 10.1016/j.jbi.2004.08.005] [Medline: 15542015]

4. de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. J Am Med Inform Assoc. 2011;18(5):557-562. [doi: 10.1136/amiajnl-2011-000150] [Medline: 21565856]
5. Torii M, Wagholikar K, Liu H. Detecting concept mentions in biomedical text using hidden Markov model: multiple concept types at once or one at a time? J Biomed Semantics. Jan 17, 2014;5(1):3. [doi: 10.1186/2041-1480-5-3] [Medline: 24438362]
6. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data. May 24, 2016;3(1):160035. [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
7. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. Feb 15, 2020;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
8. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. Presented at: Proceedings of the 2nd Clinical Natural Language Processing Workshop; Jun 7, 2019:72-78; Minneapolis, MN. [doi: 10.18653/v1/W19-1909]
9. Krishna K, Khosla S, Bigham J, Lipton ZC. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. Presented at: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); Aug 1-6, 2021:4958-4972; Online. [doi: 10.18653/v1/2021.acl-long.384]
10. Hu J, Li Z, Chen Z, Li Z, Wan X, Chang TH. Graph enhanced contrastive learning for radiology findings summarization. arXiv. Preprint posted online on Jun 8, 2022. URL: https://arxiv.org/abs/2204.00203 [Accessed 2022-04-01] [doi: 10.48550/arXiv.2204.00203]
11. Kanwal N, Rizzo G. Attention-based clinical note summarization. arXiv. Preprint posted online on Apr 18, 2021. URL: https://arxiv.org/abs/2104.08942 [Accessed 2021-04-18] [doi: 10.48550/arXiv.2104.08942]
12. Zhang Y, Zhang Y, Qi P, Manning CD, Langlotz CP. Biomedical and clinical English model packages for the Stanza Python NLP library. J Am Med Inform Assoc. Aug 13, 2021;28(9):1892-1899. [doi: 10.1093/jamia/ocab090] [Medline: 34157094]
13. Wei Q, Ji Z, Li Z, et al. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. J Am Med Inform Assoc. Jan 1, 2020;27(1):13-21. [doi: 10.1093/jamia/ocz063] [Medline: 31135882]
14. Roberts K, Shooshan SE, Rodriguez L, Abhyankar S, Kilicoglu H, Demner-Fushman D. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. J Biomed Inform. Dec 2015;58 Suppl(Suppl):S111-S119. [doi: 10.1016/j.jbi.2015.06.010] [Medline: 26122527]
15. Yang X, Bian J, Hogan WR, Wu Y. Clinical concept extraction using transformers. J Am Med Inform Assoc. Dec 9, 2020;27(12):1935-1942. [doi: 10.1093/jamia/ocaa189] [Medline: 33120431]
16. Devlin J, Chang MW, Lee K, Toutanova K. Pre-training of deep bidirectional transformers for language understanding. Presented at: 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); Jun 2-7, 2019:4171-4186; Minneapolis, MN. [doi: 10.18653/v1/N19-1423]
17. SKTBrain/KoBERT: Korean BERT pre-trained cased (KoBERT). GitHub. 2019. URL: https://github.com/SKTBrain/KoBERT.git [Accessed 2022-05-02]
18. Pires T, Schlinger E, Garrette D. How multilingual is multilingual BERT? Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Jul 28 to Aug 2, 2019:4996-5001; Florence, Italy. [doi: 10.18653/v1/P19-1493]
19. Percha B, Pisapati K, Gao C, Schmidt H. Natural language inference for curation of structured clinical registries from unstructured text. J Am Med Inform Assoc. Dec 28, 2021;29(1):97-108. [doi: 10.1093/jamia/ocab243] [Medline: 34791282]
20. Romanov A, Shivade C. Lessons from natural language inference in the clinical domain. Presented at: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; Oct 31 to Nov 4, 2018:1586-1596; Brussels, Belgium. [doi: 10.18653/v1/D18-1187]
21. El Boukkouri H, Ferret O, Lavergne T, Noji H, Zweigenbaum P, Tsujii J. CharacterBERT: reconciling ELMo and BERT for word-level open-vocabulary representations from characters. Presented at: Proceedings of the 28th International Conference on Computational Linguistics; Dec 8-13, 2020:6903-6915; Barcelona, Spain. [doi: 10.18653/v1/2020.coling-main.609]
22. Kanakarajan KR, Kundumani B, Sankarasubbu M. BioELECTRA: pretrained biomedical text encoder using discriminators. Presented at: Proceedings of the 20th Workshop on Biomedical Language Processing; Jun 11, 2021:143-154; Online. [doi: 10.18653/v1/2021.bionlp-1.16]

23. Clark K, Luong MT, Le QV, Manning CD. Electra: pre-training text encoders as discriminators rather than generators. arXiv. Preprint posted online on Mar 23, 2020. URL: https://arxiv.org/abs/2003.10555 [Accessed 2023-09-15] [doi: 10.48550/arXiv.2003.10555]
24. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med. May 20, 2021;4(1):86. [doi: 10.1038/s41746-021-00455-y] [Medline: 34017034]
25. Zhang N, Jankowski M. Hierarchical BERT for medical document understanding. arXiv. Preprint posted online on Mar 11, 2022. URL: https://arxiv.org/abs/2204.09600 [Accessed 2023-09-15] [doi: 10.48550/arXiv.2204.09600]
26. Pampari A, Raghavan P, Liang J, Peng J. EmrQA: a large corpus for question answering on electronic medical records. Presented at: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; Oct 31 to Nov 4, 2018:2357-2368; Brussels, Belgium. [doi: 10.18653/v1/D18-1258]
27. Yue X, Gutierrez BJ, Sun H. Clinical reading comprehension: a thorough analysis of the emrAQ dataset. Presented at: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Jul 5-10, 2020:4474-4486; Online. [doi: 10.18653/v1/2020.acl-main.410]
28. Rawat BPS, Weng WH, Min SY, Raghavan P, Szolovits P. Entity-enriched neural models for clinical question answering. Presented at: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing; Jul 9, 2020:112-122; Online. [doi: 10.18653/v1/2020.bionlp-1.12]
29. Savery M, Abacha AB, Gayen S, Demner-Fushman D. Question-driven summarization of answers to consumer health questions. Sci Data. Oct 2, 2020;7(1):322. [doi: 10.1038/s41597-020-00667-z] [Medline: 33009402]
30. 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. Sep 1, 2011;18(5):552-556. [doi: 10.1136/amiajnl-2011-000203]
31. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc. 2013;20(5):806-813. [doi: 10.1136/amiajnl-2013-001628] [Medline: 23564629]
32. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inform Assoc. Jan 1, 2020;27(1):3-12. [doi: 10.1093/jamia/ocz166]
33. 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 posted online on Sep 26, 2019. URL: https://arxiv.org/abs/1909.11942 [Accessed 2023-09-15] [doi: 10.48550/arXiv.1909.11942]
34. Liu Z, Lin W, Shi Y, Zhao J. A robustly optimized BERT pre-training approach with post-training. Presented at: Proceedings of the 20th Chinese National Conference on Computational Linguistics; Aug 13-15, 2021:1218-1227; Huhhot, China.
35. Richie R, Ruiz VM, Han S, Shi L, Tsui FR. Extracting social determinants of health events with transformer-based multitask, multilabel named entity recognition. J Am Med Inform Assoc. Jul 19, 2023;30(8):1379-1388. [doi: 10.1093/jamia/ocad046] [Medline: 37002953]
36. Lybarger K, Yetisgen M, Uzuner Ö. The 2022 n2c2/UW shared task on extracting social determinants of health. J Am Med Inform Assoc. Jul 19, 2023;30(8):1367-1378. [doi: 10.1093/jamia/ocad012] [Medline: 36795066]
37. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN. Attention is all you need. Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); Dec 4-7, 2017:5998-6008; Long Beach, CA, USA.
38. Zhu Y, Kiros R, Zemel R, et al. Aligning books and movies: towards story-like visual explanations by watching movies and reading books. Presented at: 2015 IEEE International Conference on Computer Vision (ICCV); Dec 7-13, 2015:19-27; Santiago, Chile. [doi: 10.1109/ICCV.2015.11]
39. Kudo T, Richardson J. SentencePiece: a simple and language independent subword tokenizer and detokenizer for neural text processing. Presented at: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; Oct 31 to Nov 4, 2018:66-71; Brussels, Belgium. [doi: 10.18653/v1/D18-2012]
40. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Nov 3-7, 2019:3982-3992; Hong Kong, China. [doi: 10.18653/v1/D19-1410]
41. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010;17(3):229-236. [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]
42. medinfoman/multifaceted-berts: a study that verified the performance of BERT models in clinical text from various perspectives. GitHub. URL: https://github.com/medinfoman/multifaceted-berts.git [Accessed 2024-03-09]
43. Kim YM, Lee TH. Korean clinical entity recognition from diagnosis text using BERT. BMC Med Inform Decis Mak. Sep 30, 2020;20(Suppl 7):242. [doi: 10.1186/s12911-020-01241-8] [Medline: 32998724]

44. Oh SH, Kang M, Lee YH. Protected health information recognition by fine-tuning a pre-training transformer model. Healthc Inform Res. Jan 2022;28(1):16-24. [doi: 10.4258/hir.2022.28.1.16]

Abbreviations

[CLS]: classification
[UNK]: unknown
ALBERT: A Lite BERT
BERT: bidirectional encoder representations from transformers
BioBERT: BERT for Biomedical Text Mining
ELECTRA: efficiently learning an encoder that classifies token replacements accurately
emrQA: electronic medical record question answering
FFNN: feedforward neural network
GPU: graphics processing unit
HMM: hidden Markov model
IRB: institutional review board
KoBERT: Korean BERT
LLM: large language model
M-BERT: Multilingual BERT
MIMIC-III: Medical Information Mart for Intensive Care
NLP: natural language processing
RoBERTa: Robustly Optimized BERT Pretraining Approach
SOAP: Subjective, Objective, Assessment, Plan

Edited by Christian Lovis; peer-reviewed by Christina Haag, Dillon Chrimes, Maria Chatzimina; submitted 19.09.2023; final revised version received 08.07.2024; accepted 17.08.2024; published 30.10.2024

Please cite as:
Kim K, Park S, Min J, Park S, Kim JY, Eun J, Jung K, Park YE, Kim E, Lee EY, Lee J, Choi J
Multifaceted Natural Language Processing Task–Based Evaluation of Bidirectional Encoder Representations From Transformers Models for Bilingual (Korean and English) Clinical Notes: Algorithm Development and Validation
JMIR Med Inform 2024;12:e52897
URL: https://medinform.jmir.org/2024/1/e52897
doi: 10.2196/52897