Original Paper

Modified Bidirectional Encoder Representations From Transformers Extractive Summarization Model for Hospital Information Systems Based on Character-Level Tokens (AlphaBERT): Development and Performance Evaluation

Yen-Pin Chen[1,2,3], MD; Yi-Ying Chen[3], MD; Jr-Jiun Lin[3], MD; Chien-Hua Huang[3,4], MD, PhD; Feipei Lai[1,5,6], PhD

[1]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei City, Taiwan
[2]Department of Emergency Medicine, National Taiwan University Hospital Chu-Tung Branch, Hsinchu County, Taiwan
[3]Department of Emergency Medicine, National Taiwan University Hospital, Taipei City, Taiwan
[4]Department of Emergency Medicine, College of Medicine, National Taiwan University, Taipei City, Taiwan
[5]Department of Computer Science & Information Engineering, National Taiwan University, Taipei City, Taiwan
[6]Department of Electrical Engineering, National Taiwan University, Taipei City, Taiwan

Corresponding Author:
Yen-Pin Chen, MD
Graduate Institute of Biomedical Electronics and Bioinformatics
National Taiwan University
Room 410, Barry Lam Hall
No 1, Sec 4, Roosevelt Road
Taipei City,
Taiwan
Phone: 886 2 3366 3754
Email: f06945029@g.ntu.edu.tw

Abstract

Background: Doctors must care for many patients simultaneously, and it is time-consuming to find and examine all patients' medical histories. Discharge diagnoses provide hospital staff with sufficient information to enable handling multiple patients; however, the excessive amount of words in the diagnostic sentences poses problems. Deep learning may be an effective solution to overcome this problem, but the use of such a heavy model may also add another obstacle to systems with limited computing resources.

Objective: We aimed to build a diagnoses-extractive summarization model for hospital information systems and provide a service that can be operated even with limited computing resources.

Methods: We used a Bidirectional Encoder Representations from Transformers (BERT)-based structure with a two-stage training method based on 258,050 discharge diagnoses obtained from the National Taiwan University Hospital Integrated Medical Database, and the highlighted extractive summaries written by experienced doctors were labeled. The model size was reduced using a character-level token, the number of parameters was decreased from 108,523,714 to 963,496, and the model was pretrained using random mask characters in the discharge diagnoses and International Statistical Classification of Diseases and Related Health Problems sets. We then fine-tuned the model using summary labels and cleaned up the prediction results by averaging all probabilities for entire words to prevent character level–induced fragment words. Model performance was evaluated against existing models BERT, BioBERT, and Long Short-Term Memory (LSTM) using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) L score, and a questionnaire website was built to collect feedback from more doctors for each summary proposal.

Results: The area under the receiver operating characteristic curve values of the summary proposals were 0.928, 0.941, 0.899, and 0.947 for BERT, BioBERT, LSTM, and the proposed model (AlphaBERT), respectively. The ROUGE-L scores were 0.697, 0.711, 0.648, and 0.693 for BERT, BioBERT, LSTM, and AlphaBERT, respectively. The mean (SD) critique scores from doctors were 2.232 (0.832), 2.134 (0.877), 2.207 (0.844), 1.927 (0.910), and 2.126 (0.874) for reference-by-doctor labels, BERT, BioBERT, LSTM, and AlphaBERT, respectively. Based on the paired t test, there was a statistically significant difference in LSTM compared to the reference (*P*<.001), BERT (*P*=.001), BioBERT (*P*<.001), and AlphaBERT (*P*=.002), but not in the other models.

Conclusions: Use of character-level tokens in a BERT model can greatly decrease the model size without significantly reducing performance for diagnoses summarization. A well-developed deep-learning model will enhance doctors' abilities to manage patients and promote medical studies by providing the capability to use extensive unstructured free-text notes.

(JMIR Med Inform 2020;8(4):e17787) doi: 10.2196/17787

KEYWORDS

transformer; BERT; deep learning; emergency medicine; automatic summarization

Introduction

Background

Medical centers are the last line of defense for public health and are responsible for educating medical talent. The number of patients in the emergency department of such medical centers is particularly large, and these patients tend to have more severe conditions than those admitted to hospital at a lower tier. For staff, the emergency department can be an overloaded work environment [1,2]. At the beginning of the shift, a doctor must perform primary care for more than 30 patients who remain in the emergency department from less than 1 hour to more than 3 days, while simultaneously treating new arrivals from triage. The conditions of patients in the emergency department also tend to change rapidly, and the staff must be able to handle these patients under time constraints. The International Statistical Classification of Diseases and Related Health Problems (ICD) codes [3] and recent discharge diagnoses can help staff rapidly determine baseline conditions. However, in a medical center, patients may have multiple underlying diseases and several comorbidities that were previously recorded as ICD codes and discharge diagnoses in electronic health records (EHRs). Because ICD codes only reflect the disease and not the associated treatments, this lack of information limits the ability of medical staff to consider information related to a previous hospital visit. Occasionally, ICD codes are selected imprecisely and do not adequately represent the condition of the patient. Therefore, discharge diagnoses are required for staff to become familiar with a patient's condition. However, the number of words describing these details in a diagnostic sentence can vary widely. Consequently, the attending physician in the emergency department may have to read as many as 1500 words to cover the medical history of all patients under their charge. To resolve this challenge, the purpose of this study was to establish a diagnostic summary system to help hospital staff members check information on all patients more quickly.

Related Works

There are several available methods to accomplish a text summarization task, ranging from traditional natural language processing (NLP) to deep-learning language models [4-9]. The goals of previous text summarization studies in the medical field [5] included finding information related to patient care in the medical literature [5,10-13], identifying drug information [14], determining medical article topic classifications [15], and summarizing medical articles [16]. In the majority of cases, data sources for the automatic summarization task were medical articles [16] such as PubMed articles [5,11,14,15]. In recent years, EHRs have been widely adopted in several hospitals and clinics, and additional data sources such as the Medical Information Mart for Intensive Care III [17] dataset are available online for free and promote medical progress. Based on medical record research, the monitoring of several disease indicators, clinical trial recruitments, and clinical decision making, several clinical summarization systems based on EHRs have been studied [4,18-20]. However, no studies have addressed the issue of a diagnostic summary system to help hospital staff access information on all patients in their care more quickly.

Although EHRs provide useful information, the majority of this information is recorded as free text, making it challenging to analyze along with other structured data [4]. In recent years, NLP and deep-learning approaches have flourished, furnishing health care providers with a new field to promote human health. Several excellent language models are now available to help machines analyze free text. One such model is Bidirectional Encoder Representations from Transformers (BERT) [21], which is an extension of Transformer [22], and received the highest score for several NLP tasks [21,23,24].

Transformer is a state-of-the-art model, which was released to translate and improve the efficiency of Long Short-Term Memory (LSTM) [25]-based language models [22]. Similar to many deep-network models, Transformer has an encoder and a decoder. The encoder converts the input data into meaningful codes (vector or matrix), while reducing the dimension size (a major bottleneck for data analysis), and the decoder converts the code to output [26]. Taking translation as an example, the encoder converts an English sentence into a digital vector in latent space, and the decoder then converts the digital vector into a corresponding sentence in the desired language. The encoder of Transformer has an embedding model, a repeating block model with a multihead self-attention model, and a feedforward model with an architecture based on the shortcut connections concept [27] and layer normalization [22,28].

The automatic text summarization task has two branches: extractive and abstractive [29]. The extractive branch identifies keywords or sentences as summaries without changing the original document, while the abstractive branch adapts a new short sentence. The diagnosis summarizes the entire admission course, including the chief complaints and treatment course, in highly concentrated and meaningful sentences that help other staff members to quickly manage patients. Because patients in the emergency department have many underlying diseases, along with the high complexity of the conditions of individual patients, incomplete sentences, grammatical issues, and some subordinate prompts, the diagnosis obtained may not be concise. Consequently, the staff needs to include an abundance of words in their diagnoses to best represent the condition of the patient. These rich vocabularies involve not only specific disease terms but also important treatments that are delivered in the course of admission and are associated with verbose text related to diagnoses. Therefore, it is necessary to further summarize the diagnoses using an extractive summarization approach.

The extractive summarization model can be simplified to a regression problem that outputs the probability of choosing or not choosing. Taking a single character as the token unit, this problem is similar to the segmentation problem in computer vision [30,31], which outputs the class probability by pixels. A BERT-based model is the superior choice in this context since the attention weight is similar to the extraction probability [32,33] and Transformer was reported to exhibit higher performance with the language model than convolutional neural networks, recurrent neural networks, or the LSTM model [22].

BERT is a state-of-the-art language model for many NLP tasks that is pretrained with unsupervised learning, including “masked language modeling” and “next-sentence prediction.” BERT is pretrained through several corpus datasets, which are then transferred to learning through supervised data [34,35] to defeat other language models in several competitions [21,36]. The pretrained model is available [37] and can be fine-tuned for many scenarios.

Because English is not the native language in Taiwan, there are various typos and spelling errors in free-text medical records. Use of the word-level method [38], which is based on Word2vec [39,40], can result in this out-of-vocabulary obstacle. In addition, the internal structure of the word is also important and improves vector representation [41,42]. This obstacle can be overcome by adopting the character-level method [40,43,44], which uses a single character or letter as the analysis unit, or the byte-pair encoding (BPE) model, which breaks down each word into multiple subword units (ie, “word pieces”) [45]. These methods can decrease the total vocabulary and can also handle rare words, typos, and spelling errors. The word-level and BPE methods were adopted in BERT, resulting in a comprehensive and adaptable model for many types of NLP tasks.

In EHRs, medical terms, abbreviations, dates, and some count numbers for treatment are rarely found in the general corpus dataset, and will result in poor performance of the model. BioBERT, which is based on the BERT model and uses the same tokenizer, is obtained through advanced training on a biomedical corpus [46], and was considered to be well-suited to address our study aims. However, the general computing environments of some medical centers have limited capability to train or fine-tune a heavy model (involving approximately 1 billion parameters) in BERT. Therefore, replacing token units with a character-level method can further reduce the vocabulary and model size, enabling the use of the internal structures of words to avoid the out-of-vocabulary problem.

Objective

Our goal was to build a diagnoses-extractive summarization model that can run on the limited computing resources of hospital information systems with good performance. Therefore, we present AlphaBERT, a BERT-based model using the English alphabet (character-level) as the token unit. We compared the performance of AlphaBERT and the number of parameters with those of the other existing models described above.

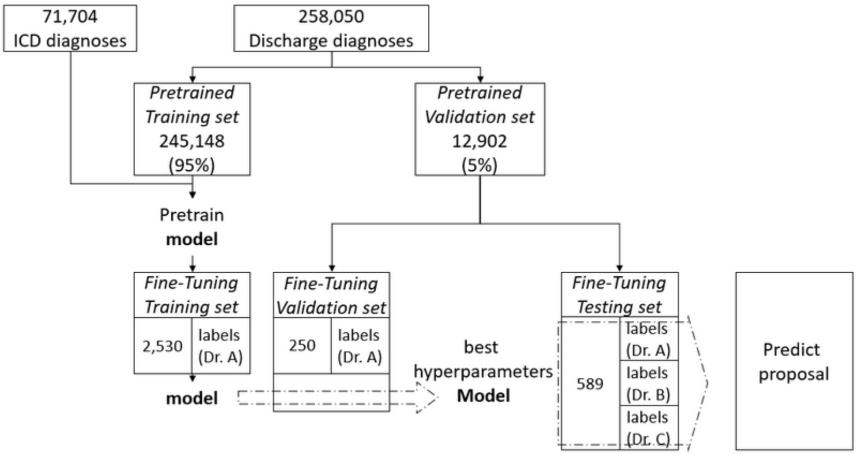
Methods

Materials

A dataset of 258,050 discharge diagnoses was obtained from the National Taiwan University Hospital Integrated Medical Database (NTUH-iMD). The discharge diagnoses originated from the following departments (in descending order): surgery, internal medicine, obstetrics and gynecology, pediatrics, oncology, orthopedic surgery, urology, otolaryngology, ophthalmology, traumatology, dentistry, neurology, family medicine, psychiatry, physical medicine and rehabilitation, dermatology, emergency medicine, geriatrics, and gerontology. This study was approved by Research Ethics Committee B, National Taiwan University Hospital (201710066RINB).

In the pretraining stage, 71,704 diagnoses collected by the ICD 10th Revision (ICD-10) [3] were also used, and the 258,050 discharge diagnoses were split into 245,148 (95.00%) as the pretrained training dataset and 12,902 (5.00%) as the pretrained validation dataset. In the fine-tuning stage, the extractive summary for supervised learning was labeled by three experienced doctors who have worked in the emergency department for more than 8 years. The fine-tuned dataset included 2530 training labels from the pretrained training dataset, and 250 validation labels and 589 testing labels from the pretrained validation dataset (Figure 1). We fed the model using 589 data entries in the fine-tuning testing set and obtained a predicted proposal for performance evaluation.

Figure 1. Pretrained validation dataset. ICD: International Statistical Classification of Diseases and Related Health Problems.



Implementation Details

The hardware used for implementation was an I7 5960x CPU, with 60 G RAM, and 2 Nvidia GTX 1080 Ti GPUs. The software used were Ubuntu 18.04 [47], Anaconda 2019.03 [48], and PyTorch 1.2.0 [49].

Label Data

We created a diagnosis-label tool to print the discharge diagnosis from the dataset in a textbox. Doctors highlighted the discharge diagnoses by selecting words that were considered to be most relevant, and the tool identified the highlighted position characters, which were labeled 1 and the others were labeled 0. For example, “1.Bladder cancer with” was labeled “0011111111111111110000” and stored in the label dataset. We encouraged doctors to skip short diagnoses, because the summarization service will be more useful for longer diagnoses. Therefore, only longer diagnoses were labeled and collected in the fine-tuning set.

Data Augmentation

In this study, the pretraining dataset was smaller than the dataset used in the pretrained model of BERT and its extensions [21,46]. Because the diagnoses included several independent diagnoses such as hypertension, cellulitis, and colon cancer, we augmented the pretraining dataset by stitching many diagnoses derived from ICD codes or NTUH-iMD. Accordingly, data augmentation was performed by selecting between 1 and 29 random diagnostic data entries from the dataset and combining them into longer and more complex diagnoses as the pretrained dataset. We set all diagnoses to a maximum of 1350 characters because of GPU memory limitations.

Because there was also a significant shortage of fine-tuning data, the same data augmentation strategy was used to extend the fine-tuning dataset. To provide greater tolerance for typos, we also randomly replaced 0.1% of the characters in the diagnoses during the fine-tuning stage.

Preprocess and Tokenization

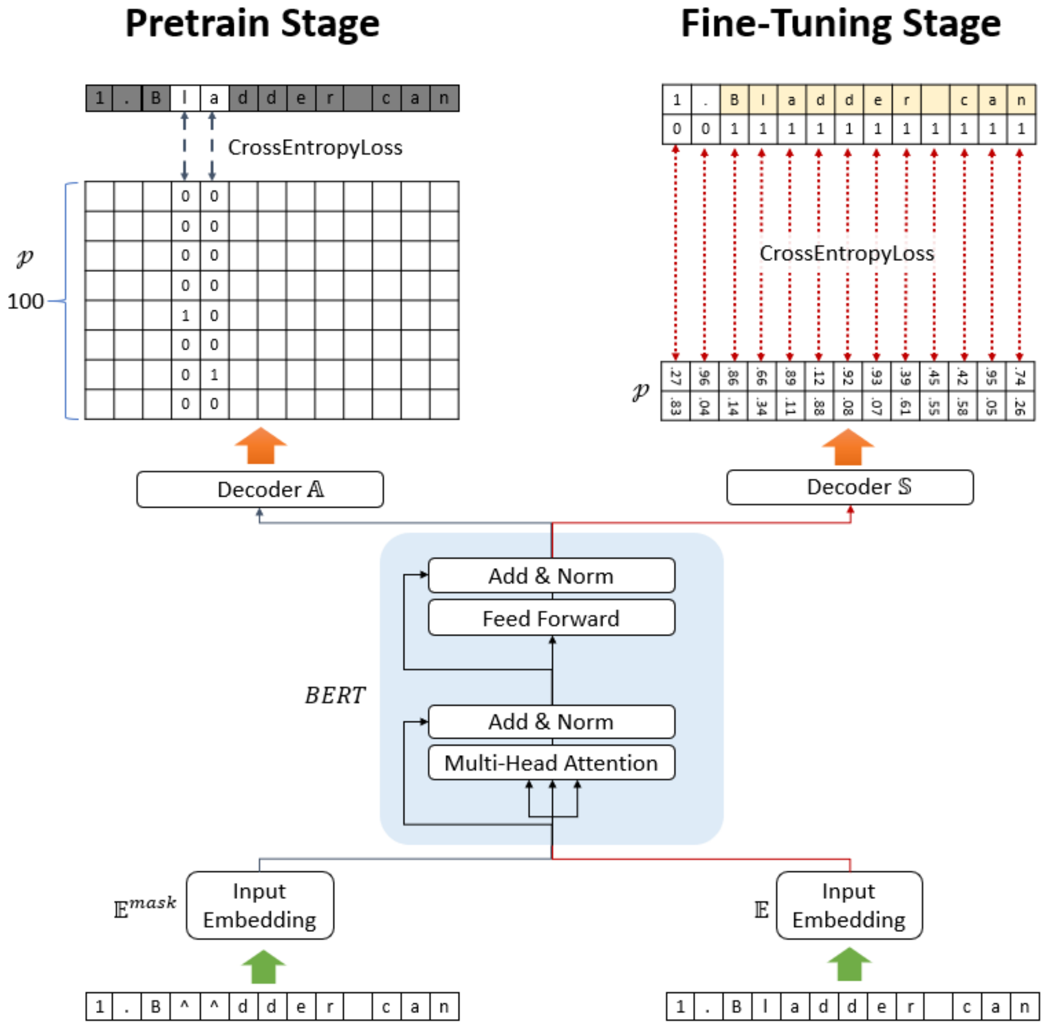
We retained only 100 symbols, including letters, numbers, and some punctuation. All free-text diagnoses were preprocessed by filters, and symbols outside of the reserved list were replaced with spaces. Original letter cases (uppercase and lowercase) were retained for analysis.

The preprocessing of diagnoses then converted the symbols (letters, numbers, and punctuation) into numbers with a one-to-one correspondence. For example, “1.Bladder cancer with” was converted to the array “14, 11, 31, 68, 57, 60, 60, 61, 74, 0, 59, 57, 70, 59, 61, 74, 0, 79, 65, 76, 64.”

Model Architecture

The architecture of AlphaBERT is based on that of BERT, and our model is based on the PyTorch adaptation released by the HuggingFace team [37] . In this study, we used a 16-layer Transformer encoder with 16 self-attention heads and a hidden size of 64. Character-level tokenizers were used as the token generator of AlphaBERT. There are 963,496 parameters in the whole model, and the symbols are represented by tokenization as one-hot encoding, corresponding to each vector with a hidden size of 64 as the token embeddings. The position embeddings (hidden size 64) are trainable vectors that correspond to the position of the symbol [21], in which the maximum length of position embeddings is set to 1350. The summation of the token embeddings and position embeddings is then used as the input embeddings (Multimedia Appendix 1) as input to AlphaBERT (Figure 2).

Figure 2. Deep-learning model architecture.



Pretraining Stage

The two-stage learning approach of BERT [21] is based on an unsupervised feature-based method, which then transfers the learning to supervised data. The unsupervised pretraining stage of BERT uses a masked language model procedure called a "cloze procedure" [21,50]. Since AlphaBERT was used as the character-level token model, and we used "^" as the "[MASK]" in BERT, we randomly selected 15% of the character sequence, 80% of which was replaced by "^," 10% was replaced with letters, and the remaining 10% was left unchanged. After the loss converged, we then masked the entire word to further pretrain our model.

Because the free-text diagnoses contained dates, chemotherapy cycles, cancer staging index, and punctuation marks, these words were nonprompted, nongeneric, and changed sequentially. Even experienced doctors cannot recover hidden dates or cycles without prompts, and therefore the letters were replaced with other letters, numbers were replaced with other numbers, and punctuation marks were replaced with other punctuation marks (but were still randomly selected to mask by "^").

In the masked language model used in this study, the BERT model was connected to a fully connected network decoder **A**, which then transformed the 64-dimensional hidden size to a 100-dimensional symbol list size corresponding to the probability *p* of each symbol. The loss function *Loss*^{mask} is the cross-entropy among the probabilities of each symbol (left side of Figure 2).

$$\mathbb{Loss}^{mask} = \sum_{c=0}^{99}\sum_{i=1}^{n} 1_i^{mask}\left[-y_{c,i}\log\left(\wp\left(\mathbb{A}\left(BERT\left(\mathbb{E}^{mask}\right)\right)\right)_{c,i}\right)\right]$$

where E^{mask} denotes the input embedding converted from masking characters, *BERT ()* is the BERT model, *A* () is the fully connected linear decoder to each preserved character, *p* is the probability function, and 1_i^{mask} denotes the i_{th} character masked.

Fine-Tuning Stage

Another fully connected network, ***S***, decoded the results of the multi-layer Transformer encoder to the predicted probability *p*. The output size of the decoder ***S*** is two-dimensional, which indicated the possibility of selection. The loss function ***Loss*** is the cross-entropy among *p* and the ground truth (right side of Figure 2).

where *S* () is the full connected linear decoder for selection.

Cleanup Method

When we evaluated our model, the probability of each word was represented by the mean probability of each character in the word. In this method, we split the characters list $C = [c_1, c_2,...c_n]$ into a list of several word sets $W = [w_1, w_2, ..., w_k]$, $k \leq n$, where the cleanup probability $p^{\wedge}_i$ of each c_i will be the average of all probabilities in w_m that contain c_i.

$$\hat{p}_i = \frac{\sum_{c_j \in w_m} p_j}{n(w_m)}\Bigg|_{c_i \in w_m}$$

where p denotes the probability after clean up, w_m denotes the sequences of characters belonging to the m_{th} word, and $n()$ is the length of the unit in the set.

BERT Models for Extractive Summarization

We also compared the state-of-the-art models and adjusted them to fit the target task. The purpose of these models was not summarization, and there is no well-presented, fine-tuned model for this purpose available. Based on the word pieces BPE method [45], all words were split into several element tokens and then the predicted result was associated with the word pieces. Accordingly, for this task, we filtered out the punctuation marks and added "[CLS]" in the head of every word (E^{head}) o represent the entire word, which prevented fragmented results.

$$\mathbb{L}\text{oss}^{head} = \sum_{c=0}^{1}\sum_{i=1}^{n} 1_i^{head}\left[-y_{c,i}\log\left(p\left(\mathbb{S}\left(BERT\left(\mathbb{E}^{head}\right)\right)\right)_{c,i}\right)\right]$$

Where E^{head} denotes the input embedding converted from a word (with head) and 1_i^{head} denotes that the i_{th} character is a head token.

LSTM Model for Extractive Summarization

We also used the LSTM model [23,25] for this summarization task. To achieve effective comparison with our model, we pretrained the input embedding using Word2vec [39] and adopted a 9-layer bidirectional LSTM with 899,841 parameters, which was very similar to our model.

$$\mathbb{L}\text{oss} = \sum_{c=0}^{1}\sum_{i=1}^{n} -y_{c,i}\log\left(p\left(\mathbb{S}\left(LSTM\left(\mathbb{E}\right)\right)\right)_{c,i}\right)$$

Hyperparameters

We used Adam optimization [51] with a learning rate of 1×10^{-5} in the warmup phase [27,52,53], and then switched to a rate of 1×10^{-4} and a minibatch size of 2. The hyperparameter used in this study was the threshold to the character-level probability of selection, which was chosen using a receiver operating characteristic (ROC) curve and *F1* statistic counting from the fine-tuning validation set (Multimedia Appendix 2).

Measurement

We measured the performance of the various models using the ROC curve, an *F1* statistic, and the *F1* statistic of Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [54]. To maintain measurement consistency, we filtered out all punctuation in the predicted proposals, counted the results at the word level, and collected physicians' feedback for each model. A questionnaire website was established in which the original diagnoses were randomly selected and displayed in the first part, and the ground truth summary proposal determined by testing labels and proposals predicted by models were displayed in the second part under random sorting. We recruited 14 experienced physicians for this purpose, including the chief resident, 10 attending physicians of the emergency department at the medical center, one emergency department attending physician at the regional hospital, and two emergency attending physicians at the district hospital. They entered a score of 0-3 for each proposal, in which 0 represented "nonsensical" and 3 represented "good."

Statistical Analysis

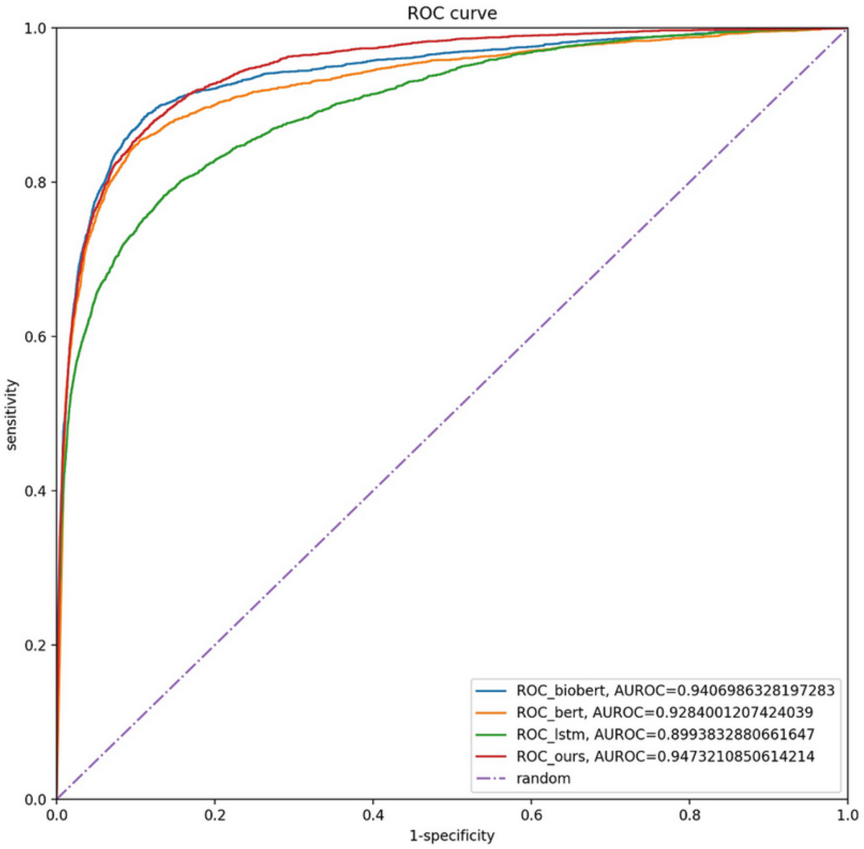
Data were analyzed using the statistical package RStudio (version 1.2.5019) based on R (version 3.6.1; R Foundation for Statistical Computing, Vienna, Austria). For group comparisons, we performed the pairwise paired *t* test on the dependent variables of the physician scores and set the significance threshold level to $P<.05$.

Results

The discharge diagnoses dataset included 57,960 lowercase English words. The maximum number of words in a diagnosis was 654 (3654 characters), with a mean of 55 (SD 51) words corresponding to 355 (SD 318) characters. In the fine-tuning dataset, the mean number of words in the diagnoses and summary were 78 (SD 56) and 12 (SD 7), respectively. The retention ratio [55] (ie, words in the summary divided by words in the diagnoses) was 12 out of 78 words (15%). The fine-tuning testing set included 138 diagnoses with incorrect words, and a total of 183 incorrect words were counted manually by two attending physicians, including 153 misspellings, 13 typos, 14 inappropriate words, and 3 repeated words.

Our proposed model, AlphaBERT, demonstrated the highest performance among all compared models with an area under the ROC curve (AUROC) of 0.947, and the LSTM demonstrated the worst performance with an AUROC of 0.899 (Figure 3).

Figure 3. Model receiver operating characteristic (ROC) curves.



BioBERT achieved the highest ROUGE scores (Table 1). BERT and the proposed model were in the intermediate range, with the lowest scores obtained with the LSTM. In addition, the ROUGE score was the highest for reference Doctor A and was the lowest for Doctor C (Table 1). When there were incorrect words in the input diagnoses, the performance of all models deteriorated (Table 2).

We collected 246 critical scores from the 14 doctors that responded to the questionnaire. Statistically significant differences (based on the paired *t* test) were detected within the LSTM compared to the reference, BERT, BioBERT, and our proposed model, but not with respect to the other models (Table 3).

We built the service on a website [56] using a server with only one CPU (no GPU) on the Microsoft Azure platform to provide a diagnoses-extractive summarization service. Editorial suggestions are also available on the website to gather user feedback and to continue to improve the model. The source code is available on GitHub [57]. The service is currently being integrated into the hospital information system to enhance the capabilities of hospital staff.

Table 1. Model parameters and ROUGE[a] F1 results.

Model	Dr A (n=250)	Dr B (n=248)	Dr C (n=91)	Mean *F1* value
BERT[b] (108,523,714 parameters)				
ROUGE-1[c]	0.761	0.693	0.648	0.715
ROUGE-2[d]	0.612	0.513	0.473	0.549
ROUGE-L[e]	0.748	0.671	0.627	0.697
BioBERT[f] (108,523,714 parameters)				
ROUGE-1	0.788	0.697	0.647	0.728
ROUGE-2	0.642	0.523	0.464	0.565
ROUGE-L	0.773	0.678	0.629	0.711
LSTM[g] (899,841 parameters)				
ROUGE-1	0.701	0.647	0.618	0.666
ROUGE-2	0.531	0.468	0.459	0.494
ROUGE-L	0.684	0.629	0.602	0.648
Proposed model (963,496 parameters)				
ROUGE-1	0.769	0.678	0.647	0.712
ROUGE-2	0.610	0.482	0.463	0.533
ROUGE-L	0.751	0.656	0.632	0.693

[a] ROUGE: Recall-Oriented Understudy for Gisting Evaluation.
[b] BERT: Bidirectional Encoder Representations from Transformers.
[c] ROUGE-1: Recall-Oriented Understudy for Gisting Evaluation with unigram overlap.
[d] ROUGE-2: Recall-Oriented Understudy for Gisting Evaluation with bigram overlap.
[e] ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation for the longest common subsequence (n) representing the number of reference labels.
[f] BioBERT: Bidirectional Encoder Representations from Transformers trained on a biomedical corpus.
[g] LSTM: Long Short-Term Memory.

Table 2. ROUGE[a] F1 results of diagnoses with incorrect words.

ROUGE-L[b]	BERT[c]	BioBERT[d]	LSTM[e]	Proposed Model
Diagnoses without error words (n=451)[f]	0.704	0.717	0.651	0.698
Diagnoses with incorrect words (n=138)	0.676	0.692	0.640	0.674

[a] ROUGE: Recall-Oriented Understudy for Gisting Evaluation.
[b] ROUGE-L: ROUGE for the longest common subsequence.
[c] BERT: Bidirectional Encoder Representations from Transformers.
[d] BioBERT: Bidirectional Encoder Representations from Transformers trained on a biomedical corpus.
[e] LSTM: Long Short-Term Memory.
[f] n represents the number of reference labels.

Table 3. Critique scores of models from doctors (N=246).

Model	Score, mean (SD)	*P* value			
		BERT[a]	BioBERT[b]	LSTM[c]	Proposed Model
Reference	2.232 (0.832)	.11	.66	<.001	.10
BERT	2.134 (0.877)		.10	.001	.89
BioBERT	2.207 (0.844)			<.001	.19
LSTM	1.927 (0.910)				.002
Proposed	2.126 (0.874)				

[a] BERT: Bidirectional Encoder Representations from Transformers.

[b] BioBERT: Bidirectional Encoder Representations from Transformers trained on a biomedical corpus.

[c] LSTM: Long Short-Term Memory.

Discussion

Principal Findings

AlphaBERT effectively performed the extractive summarization task on medical clinic notes and decreased the model size compared to BERT, reducing the number of parameters from 108,523,714 to 963,496 using a character-level tokenizer. AlphaBERT showed similar performance to BERT and BioBERT in this extractive summarization task. In spite of the heavy model, both BERT and BioBERT were demonstrated to be excellent models and well-suited for several tasks (including the primary task of this study) with small adjustments. For convenience, the model can be used in a straightforward manner to rapidly build new apps in the medical field. Because of the well pretrained NLP feature extraction model, a small label dataset (the fine-tuning training set includes only 2530 cases) is sufficient for supervised learning and achieving the goal.

In this study, we obtained high ROUGE *F1* scores for all models. In general summarization studies, the ROUGE *F1* score was typically less than 0.40 [6-9], whereas we achieved a score of 0.71, which corresponds with a higher retention ratio (15%) for this task than the corpus of other summarization tasks such as the CNN/Daily Mail Corpus (approximately 7%) [7]. Since the diagnosis can be considered as a summary of admission records, a higher retention rate is reasonable; however, for emergencies, the diagnosis will contain too many redundant words in some cases.

The ICD-10 is a well-classified system with more than 70,000 codes, but is often too simple to fully capture the complex context of a patient's record. The treatments during the patient's previous hospitalization are also important to consider, and are often recorded as a free-text diagnosis when the patient has revisited a hospital under critical status. For example, if a patient has cancer, the previous chemotherapy course is important information when the patient is seriously ill in the emergency department. Furthermore, it is difficult for doctors to accurately find the correct codes; thus, it is insufficient to represent a patient's condition by simply obtaining the ICD-10 code from the EHR. However, the ICD-10 codes can be used to extend the pretrained training set by random stitching.

Combining a random number of diagnoses not only extends the training dataset but also improves the performance of the model. The average number of characters in a diagnosis was 355, but the range was larger (SD 318). In the absence of augmentation, the position embeddings and self-attention heads trained more in the front and demonstrated poorer performance in the back. Augmentation combines several diagnoses to lengthen the input embeddings, which can train the self-attention heads to consider all 1350 characters equally.

In the prediction phase, we obtained the probability of each character. Since a word is split into a sequence of characters, the result is fragmented, and only some characters in a word were selected by prediction. This results in a nonsense phrase and produces poor results. Accordingly, we proposed a cleanup method that selects the entire word based on the probability of all characters being present in the word. This concept is derived from the segmentation task in computer vision in which each pixel has the possibility of classifying and causing the predictions to not continue. In the field of computer vision, contour-based superpixels are chosen, and all superpixels are selected by a majority vote [31]. In this study, the average probability of an entire word represents the probability of each character and results in either the entire word being selected or none at all.

Since the summarization task is subjective, properly evaluating the performance of the model is a relevant consideration. Lack of adequate medical labels is an important issue, because labels from qualified physicians are rare and difficult to collect. Although the ROUGE score [54] is widely used in this field, it is evaluated by the same doctors' labels and even by separate split sets.

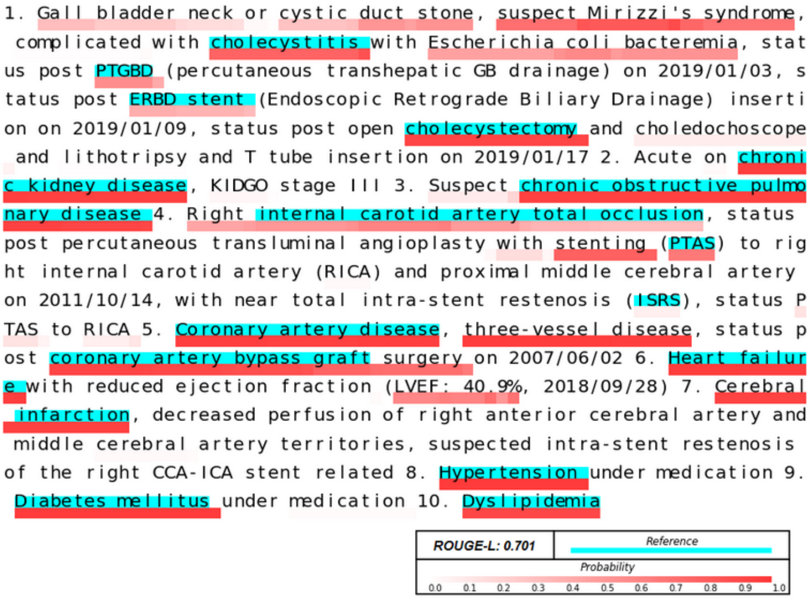
Owing to the lack of doctors who are capable of labeling the reference summaries, all of the models evaluated in this study were limited to being fine-tuned by Doctor A's labels. We were able to shuffle and randomly split the three doctors' labels to training, validation, and testing sets, but we did not have reference labels from other doctors to confirm whether individual variation exists. Even when using the three doctors' labels, this problem would occur when gathering another doctor's labels.

To confirm the differences from other doctors, the models were fine-tuned using only one doctor's knowledge, with the others' used as a test set. The results revealed a difference according to the ROUGE scores (Table 1) from the three doctors. The model had a poor ROUGE score on the label references for Doctor C, implying that summarization is a highly subjective task. Certain words are important for some doctors, but not for others, even among doctors in the same medical field who have similar interpretation processes. Therefore, it was very easy to overfit the model with the summarization task. BioBERT had the most accurate prediction result, but the associated overfitting was also more severe.

We established a website for doctors to easily critique the performance within label references and the predictions from the models to further objectively evaluate the performance of the model and the reference labels from doctors. We used a double-blind method to collect scores, and the system randomly chose a diagnosis and displayed corresponding summary proposals by random ordering. The critical reviewer was therefore blinded to the method used for each prediction. We obtained similar results to the ROUGE scores from this analysis. Moreover, the LSTM was consistently the lowest-performing model, whereas manually labeled references achieved the highest average score, followed by BioBERT.

Although the performance of AlphaBERT was not optimum, there was nevertheless no statistically significant difference between the performances of BERT, BioBERT, and AlphaBERT. The advantage of AlphaBERT is the character-level prediction probability and its one-to-one correspondence with the original document. The predicted keywords can be highlighted directly on the original document and can be easily edited by users. For example, although AlphaBERT's predicted proposal had a ROUGE-L score of 0.701, it makes sense to recognize important words, which is perhaps more informative than a doctor's reference label (Figure 4). In some cases, our proposed method could predict more information about the disease and related treatments, whereas in other cases some diseases were lost (eg, pneumonia, hypertension, and respiratory failure), and in other cases the formal medical term was predicted but the reference label was an abbreviation (Multimedia Appendix 3). This variation also reflects the subjectivity of the summary task.

Figure 4. Illustration of the performance of AlphaBERT.



1. Gall bladder neck or cystic duct stone, suspect Mirizzi's syndrome,
complicated with cholecystitis with Escherichia coli bacteremia, stat
us post PTGBD (percutaneous transhepatic GB drainage) on 2019/01/03, s
tatus post ERBD stent (Endoscopic Retrograde Biliary Drainage) inserti
on on 2019/01/09, status post open cholecystectomy and choledochoscope
and lithotripsy and T tube insertion on 2019/01/17 2. Acute on chroni
c kidney disease, KIDGO stage III 3. Suspect chronic obstructive pulmo
nary disease 4. Right internal carotid artery total occlusion, status
post percutaneous transluminal angioplasty with stenting (PTAS) to rig
ht internal carotid artery (RICA) and proximal middle cerebral artery
on 2011/10/14, with near total intra-stent restenosis (ISRS), status P
TAS to RICA 5. Coronary artery disease, three-vessel disease, status p
ost coronary artery bypass graft surgery on 2007/06/02 6. Heart failur
e with reduced ejection fraction (LVEF: 40.9%, 2018/09/28) 7. Cerebral
infarction, decreased perfusion of right anterior cerebral artery and
middle cerebral artery territories, suspected intra-stent restenosis
of the right CCA-ICA stent related 8. Hypertension under medication 9.
Diabetes mellitus under medication 10. Dyslipidemia

Limitations

Due to the subjective nature of the text summarization task, the predicted summary results may lose some information that may be of relevance. The proposed model helps hospital staff to quickly view information for a large number of patients at the beginning of a shift; however, they will still need to read all of the collected information from the EHRs during ward rounds.

Typos and misspellings remain a problem in NLP. However, the character-level and word pieces BPE method can not only reduce the vocabulary but can also handle typos effectively to maintain noninferior results (Multimedia Appendix 4). Although automatic spelling correction may be a solution to this problem, we have not included this feature in our proposed method because we are confident in the robust error tolerance of the character-level and BPE method.

This was a pilot study in the medical text summarization field based on the deep-learning method. We plan to establish a website that offers this service and provides a way to edit suggestions and feedback to collect volunteer labels and resolve personal variability in the near future.

Conclusions

AlphaBERT, using character-level tokens in a BERT-based model, can greatly decrease model size without significantly reducing performance for text summarization tasks. The proposed model will provide a method to further extract the unstructured free-text portions in EHRs to obtain an abundance of health data. As we enter the forefront of the artificial intelligence era, NLP deep-learning models are well under development. In our model, all medical free-text data can be transformed into meaningful embeddings, which will enhance medical studies and strengthen doctors' capabilities.

Acknowledgments

We would like to thank the Ministry of Science and Technology, Taiwan, for financially supporting this research (grant MOST 109-2634-F-002-029). We would also like to thank Yun-Nung Chen for providing useful comments on this work and Hugging face for providing several excellent deep-learning codings. We are grateful to GitHub for providing the code repository used for AlphaBERT.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Input embedding.
[PNG File , 17 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Flowchart to determine the hyperparameters and measure the model's performance.
[PNG File , 35 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Error statistics (strong and weak).
[PDF File (Adobe PDF File), 409 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Error statistics (typos, misspellings, or incorrect words).
[PDF File (Adobe PDF File), 703 KB-Multimedia Appendix 4]

References

1. Hsu C, Liang L, Chang Y, Juang W. Emergency department overcrowding: Quality improvement in a Taiwan Medical Center. J Formos Med Assoc 2019 Jan;118(1):186-193 [FREE Full text] [doi: 10.1016/j.jfma.2018.03.008] [Medline: 29665984]
2. Lin C, Liang H, Han C, Chen L, Hsieh C. Professional resilience among nurses working in an overcrowded emergency department in Taiwan. Int Emerg Nurs 2019 Jan;42:44-50. [doi: 10.1016/j.ienj.2018.05.005] [Medline: 29954706]
3. World Health Organization. ICD-10 Version:2019 URL: https://icd.who.int/browse10/2019/en [accessed 2016-01-01]
4. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc 2016 Dec;23(5):1007-1015 [FREE Full text] [doi: 10.1093/jamia/ocv180] [Medline: 26911811]
5. Workman TE, Fiszman M, Hurdle JF. Text summarization as a decision support aid. BMC Med Inform Decis Mak 2012 May 23;12:41 [FREE Full text] [doi: 10.1186/1472-6947-12-41] [Medline: 22621674]
6. Gigioli P, Sagar N, Rao A, Voyles J. Domain-Aware Abstractive Text Summarization for Medical Documents. 2018 Presented at: IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 3-6, 2018; Madrid, Spain URL: https://ieeexplore.ieee.org/document/8621539 [doi: 10.1109/BIBM.2018.8621539]
7. Nallapati R, Zhou B, Gulcehre C. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In: Association for Computational Linguistics.: Association for Computational Linguistics; 2016 Aug Presented at: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning; August 2016; Berlin, Germany URL: https://www.aclweb.org/anthology/K16-1028/ [doi: 10.18653/v1/K16-1028]
8. See A, Liu P. Get To The Point: Summarization with Pointer-Generator Networks. In: Association for Computational Linguistics. 2017 Jul Presented at: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); July 2017; Vancouver, Canada p. 1073-1083 URL: https://www.aclweb.org/anthology/P17-1099 [doi: 10.18653/v1/P17-1099]
9. Zhou Q, Yang N, Wei F, Huang S, Zhou M. Neural Document Summarization by Jointly Learning to Score and Select Sentences. In: Association for Computational Linguistics.: Association for Computational Linguistics; 2018 Jul Presented at: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2018/July; Melbourne, Australia p. 654-663 URL: https://www.aclweb.org/anthology/P18-1061 [doi: 10.18653/v1/p18-1061]
10. Elhadad N, McKeown K, Kaufman D, Jordan D. Facilitating physicians' access to information via tailored text summarization. AMIA Annu Symp Proc 2005:226-230 [FREE Full text] [Medline: 16779035]

11. Niu Y, Zhu X, Hirst G. Using outcome polarity in sentence extraction for medical question-answering. AMIA Annu Symp Proc 2006:599-603 [FREE Full text] [Medline: 17238411]
12. Sarker A, Mollá D, Paris C. Extractive summarisation of medical documents using domain knowledge and corpus statistics. Australas Med J 2012;5(9):478-481 [FREE Full text] [doi: 10.4066/AMJ.2012.1361] [Medline: 23115581]
13. Ranjan H, Agarwal S, Prakash A, Saha S. Automatic labelling of important terms and phrases from medical discussions. In: IEEE.: IEEE; 2017 Nov 03 Presented at: 2017 Conference on Information and Communication Technology (CICT); November 3-5, 2017; Gwalior, India URL: https://ieeexplore.ieee.org/document/8340644 [doi: 10.1109/INFOCOMTECH.2017.8340644]
14. Fiszman M, Rindflesch TC, Kilicoglu H. Summarizing drug information in Medline citations. AMIA Annu Symp Proc 2006:254-258 [FREE Full text] [Medline: 17238342]
15. Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC. Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation. J Biomed Inform 2009 Oct;42(5):801-813 [FREE Full text] [doi: 10.1016/j.jbi.2008.10.002] [Medline: 19022398]
16. Sarkar K, Nasipuri M. Using Machine Learning for Medical Document Summarization. Int J Database Theor Appl 2011 Mar;4(1):31-48 [FREE Full text]
17. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3(1):160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
18. Goldstein A, Shahar Y. An automated knowledge-based textual summarization system for longitudinal, multivariate clinical data. J Biomed Inform 2016 Jun;61:159-175 [FREE Full text] [doi: 10.1016/j.jbi.2016.03.022] [Medline: 27039119]
19. Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Ohe K. TEXT2TABLE: Medical Text Summarization System Based on Named Entity Recognition and Modality Identification. In: Association for Computational Linguistics.: Association for Computational Linguistics; 2009 Jun Presented at: Proceedings of the BioNLP 2009 Workshop; June 2009; Boulder, Colorado p. 185-192 URL: https://www.aclweb.org/anthology/W09-1324/ [doi: 10.3115/1572364.1572390]
20. Davis MF, Sriram S, Bush WS, Denny JC, Haines JL. Automated extraction of clinical traits of multiple sclerosis in electronic medical records. J Am Med Inform Assoc 2013 Dec;20(e2):e334-e340. [doi: 10.1136/amiajnl-2013-001999] [Medline: 24148554]
21. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Association for Computational Linguistics.: Association for Computational Linguistics; 2019 Jun 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); June 2019; Minneapolis, Minnesota p. 4171-4186 URL: https://www.aclweb.org/anthology/N19-1423/ [doi: 10.18653/v1/N19-1423]
22. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A. Attention Is All You Need. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc; Dec 2017:6000-6010.
23. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K. Deep contextualized word representations. 2018 Presented at: 2018 Conference of the North American Chapter for Computational Linguistics (NAACL); June 1-6, 2018; New Orleans, Louisiana. [doi: 10.18653/v1/n18-1202]
24. Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 10,000+ questions for machine comprehension of text. : Association for Computational Linguistics; 2016 Presented at: Proceedings of the Conference on Empirical Methods in Natural Language Processing; November 2016; Austin, Texas p. 2383-2392. [doi: 10.18653/v1/d16-1264]
25. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation 1997 Nov;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735]
26. Kingma D, Welling M. Auto-encoding variational bayes. 2013. URL: https://arxiv.org/abs/1312.6114 [accessed 2013-12-20]
27. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 Presented at: Proceedings of the IEEE conference on computer vision pattern recognition; June 26-July 1, 2016; Las Vegas, NV. [doi: 10.1109/cvpr.2016.90]
28. Ba J, Kiros J, Hinton G. arxiv. 2016 Jul 21. Layer normalization URL: https://arxiv.org/abs/1607.06450 [accessed 2016-07-21]
29. Hovy E, Chinyew L. Automated text summarization and the SUMMARIST system. In: Association for Computational Linguistics.: Association for Computational Linguistics; 1998 Oct 13 Presented at: TIPSTER '98: Proceedings of a workshop on held at Baltimore, Maryland; October 13-15, 1998; Baltimore, Maryland p. 197-214 URL: https://www.aclweb.org/anthology/W97-0704 [doi: 10.3115/1119089.1119121]
30. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. : Springer; 2015 Presented at: International Conference on Medical image computing and computer-assisted intervention; October 2015; Munich, Germany p. 234-241. [doi: 10.1007/978-3-319-24574-4_28]
31. Caelles S, Maninis K, Pont-Tuset J, Leal-Taixé L, Cremers D, Van GL. One-shot video object segmentation. 2017 Presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; July 21-26, 2017; Honolulu, Hawaii. [doi: 10.1109/cvpr.2017.565]
32. Wang Y, Lee H. Learning to encode text as human-readable summaries using generative adversarial networks. In: Association for Computational Linguistics. 2018 Presented at: 2018 Conference on Empirical Methods in Natural Language Processing;

October-November, 2018; Brussels, Belgium URL: https://www.aclweb.org/anthology/D18-1451/ [doi: 10.18653/v1/d18-1451]

33. Rush A, Chopra S, Weston J. A neural attention model for abstractive sentence summarization. In: Association for Computational Linguistics. 2015 Presented at: 2015 Conference on Empirical Methods in Natural Language Processing; September 2015; Lisbon, Portugal URL: https://www.aclweb.org/anthology/D15-1044/ [doi: 10.18653/v1/d15-1044]
34. Conneau A, Kiela D, Schwenk H, Barrault L, Bordes AJ. Supervised learning of universal sentence representations from natural language inference data. In: Association for Computational Linguistics. 2017 Presented at: 2017 Conference on Empirical Methods in Natural Language Processing; September 2017; Copenhagen, Denmark URL: https://arxiv.org/pdf/1705.02364.pdf [doi: 10.18653/v1/d17-1070]
35. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? : MIT Press; 2014 Dec Presented at: NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems; December 8-13, 2014; Montreal, Canada p. 3320-3328. [doi: 10.5555/2969033.2969197]
36. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R. arXiv. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding URL: https://arxiv.org/abs/1906.08237 [accessed 2019-06-19]
37. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A. Huggingface's transformers: State-of-the-art natural language processing. URL: https://huggingface.co/ [accessed 2019-01-01]
38. Brown P, Desouza P, Mercer R, Pietra V. Class-based n-gram models of natural language. Comput Ling 1992:467-480 [FREE Full text]
39. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. : Curran Associates Inc; 2013 Presented at: NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems; December 5-10, 2013; Lake Tahoe, Nevada p. 3111-3119. [doi: 10.5555/2999792.2999959]
40. Le H, Cerisara C, Denis A. arxiv. Do Convolutional Networks need to be Deep for Text Classification? URL: https://arxiv.org/abs/1707.04108 [accessed 2017-07-13]
41. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. Trans Assoc Comput Ling 2017 Dec;5:135-146. [doi: 10.1162/tacl_a_00051]
42. Xenouleas S, Malakasiotis P, Apidianaki M. SUM-QE: a BERT-based Summary Quality Estimation Model. In: Association for Computational Linguistics.: Association for Computational Linguistics; 2019 Nov 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); November, 2019; Hong Kong, China p. 6005-6011 URL: https://www.aclweb.org/anthology/D19-1618/ [doi: 10.18653/v1/D19-1618]
43. Al-Rfou R, Choe D, Constant N, Guo M, Jones L. Character-Level Language Modeling with Deeper Self-Attention. In: AAAI. 2019 Jul 17 Presented at: Character-level language modeling with deeper self-attention; 2019; Proceedings of the AAAI Conference on Artificial Intelligence p. 3159-3166. [doi: 10.1609/aaai.v33i01.33013159]
44. Zhang X, Zhao J, Lecun Y. Character-level Convolutional Networks for Text Classification. 2015 Presented at: Advances in Neural Information Processing Systems 28 (NIPS 2015); December 7-12, 2015; Montreal, Canada.
45. Wu Y, Schuster M, Chen Z, Le Q, Norouzi M, Macherey W. Google's neural machine translation system: Bridging the gap between human and machine translation. Trans Assoc Comput Ling 2017 Oct;5:339-351 [FREE Full text]
46. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
47. Ubuntu. Download Ubuntu Desktop URL: https://ubuntu.com/download/desktop [accessed 2019-01-01]
48. Anaconda. Solutions for Data Science Practitioners and Enterprise Machine Learning URL: https://www.anaconda.com/ [accessed 2019-01-01]
49. PyTorch. From Research to Production URL: https://pytorch.org/ [accessed 2019-01-01]
50. Taylor WL. "Cloze Procedure": A New Tool for Measuring Readability. Journalism Quart 1953 Sep 01;30(4):415-433. [doi: 10.1177/107769905303000401]
51. Kingma D. Adam: A Method for Stochastic Optimization. Adam; 2015 Presented at: International Conference for Learning Representations (ICLR) 2015; 2015; San Diego, California p. A URL: https://arxiv.org/abs/1412.6980
52. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI.: AAAI Press; 2017 Feb Presented at: Thirty-First AAAI Conference on Artificial Intelligence; February 2017; San Francisco, California p. 4278-4284.
53. Goyal P, Dollár P, Girshick R, Noordhuis P, Wesolowski L, Kyrola A. arvix. Accurate, large minibatch sgd: Training imagenet in 1 hour URL: https://arxiv.org/abs/1706.02677 [accessed 2018-04-30]
54. Lin C. ROUGE: A Package for Automatic Evaluation of Summaries. In: Association for Computational Linguistics. 2004 Jul Presented at: Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004; July 2004; Barcelona, Spain p. 74-81 URL: https://www.aclweb.org/anthology/W04-1013
55. Mitchell-Box K, Braun KL. Fathers' thoughts on breastfeeding and implications for a theory-based intervention. J Obstet Gynecol Neonatal Nurs 2012;41(6):E41-E50. [doi: 10.1111/j.1552-6909.2012.01399.x] [Medline: 22861175]

56. Chen YP. Azure. URL: http://diagnosislabelevaluateweb.azurewebsites.net/Extract [accessed 2020-01-13]
57. Chen YP. Github. AlphaBERT URL: https://github.com/wicebing/AlphaBERT.git [accessed 2020-04-10]

Abbreviations

AUROCs: Area Under the Receiver Operating Characteristics
BERT: Bidirectional Encoder Representations from Transformers
BPE: byte-pair encoding
EHR: electronic health record
ICD-10: International Statistical Classification of Diseases and Related Health Problems 10th Revision
LSTM: long short-term memory
NLP: natural language processing
NTUH-iMD: National Taiwan University Hospital Integrated Medical Database
ROC: receiver operating characteristic
ROUGE: Recall-Oriented Understudy for Gisting Evaluation

Edited by G Eysenbach; submitted 13.01.20; peer-reviewed by G Mayer, T Muto, S Ma, HH Rau; comments to author 06.02.20; revised version received 05.03.20; accepted 10.04.20; published 29.04.20

Please cite as:
Chen YP, Chen YY, Lin JJ, Huang CH, Lai F
Modified Bidirectional Encoder Representations From Transformers Extractive Summarization Model for Hospital Information Systems Based on Character-Level Tokens (AlphaBERT): Development and Performance Evaluation
JMIR Med Inform 2020;8(4):e17787
URL: http://medinform.jmir.org/2020/4/e17787/
doi: 10.2196/17787
PMID: 32347806