Efficient Clinical NLP via Tokenizer Extension and Knowledge Distillation

M. C. Neluhena
Department of Computer Science and Engineering
University of Moratuwa
malindun.21@cse.mrt.ac.lk

***Abstract*—Pretrained language models (PLMs) such as BERT have achieved state-of-the-art performance in clinical natural language processing (NLP), particularly in medical entity recognition (MER). However, their large size and computational requirements hinder deployment in real-world healthcare systems. DistilBERT provides a lightweight alternative but struggles with specialized medical terminology. Domain-adapted models like BioClinicalBERT achieve superior accuracy but are computationally heavy.**

This paper presents a three-step approach to produce an efficient, domain-aware DistilBERT model: first, the tokenizer is extended with high-frequency clinical terms to reduce semantic fragmentation; second, knowledge distillation (KD) transfers domain expertise from BioClinicalBERT to the student model; and finally, the distilled model is fine-tuned on the MACCROBAT dataset for medical entity recognition. Experiments on the MACCROBAT dataset show that the proposed approach improves MER accuracy over baseline DistilBERT, approaching BioClinicalBERT performance while retaining efficiency.

***Index Terms*—Clinical NLP, Medical Entity Recognition, Knowledge Distillation, Tokenization, Transformer Models**

I. Introduction

Electronic health records (EHRs) contain vast amounts of unstructured text describing patient histories, diagnoses, and treatments. Extracting structured information from this text is crucial for clinical decision support, automated coding, and disease surveillance.

Transformer-based PLMs such as BERT [1] achieve strong performance in NLP tasks. However, clinical text contains dense domain-specific terminology and abbreviations (e.g., "CHF" for chronic heart failure), posing challenges for general-domain models. While DistilBERT [2] reduces computational cost, it lacks domain knowledge. Domain-adapted models like BioClinicalBERT [3] perform better but are resource-intensive.

We propose a three-step approach to build an efficient, domain-aware DistilBERT:

1) **Tokenizer extension:** Augmenting DistilBERT's vocabulary with frequent clinical terms.
2) **Knowledge distillation:** Transferring domain knowledge from BioClinicalBERT to the student model with the extended tokenizer.
3) **Fine-tuning:** Training the adapted model on the MACCROBAT dataset for named entity recognition.

Our contributions are:

- A lightweight clinical NLP model combining tokenizer extension and knowledge distillation.
- Evaluation on the MACCROBAT dataset, showing improved accuracy with minimal overhead.
- Analysis of trade-offs between model size, speed, and accuracy.

II. Related Work

A. Domain-Adaptive Pretraining in Clinical NLP

Domain-adaptive pretraining (DAPT) has been widely adopted to enhance the performance of pretrained language models (PLMs) on biomedical and clinical text. Models such as BioBERT [4] and BioClinicalBERT [3] continue pretraining on domain-specific corpora such as PubMed abstracts, MIMIC-III, and clinical notes, resulting in significant improvements of 5–8% in F1 score over general-domain models like BERT. DAPT enables models to internalize domain-specific linguistic and semantic patterns, but it is computationally expensive, often requiring hundreds of GPU hours and access to large-scale, sensitive clinical data. Moreover, continuously retraining large PLMs for every sub-domain or institution is impractical, motivating the need for more parameter-efficient domain adaptation methods.

B. Lightweight and Compressed Models

To address the computational inefficiency of large PLMs, several lightweight and compressed alternatives have been proposed. Models such as DistilBERT [2], TinyBERT [5], and MobileBERT [6] employ knowledge distillation, parameter sharing, and architecture optimization to reduce model size and inference latency by 40–60% while retaining most of the performance of their teacher models. However, these general-domain compressed models often struggle with biomedical terminology, as their vocabularies and internal representations are biased toward general English. In clinical NLP tasks such as named entity recognition (NER) or relation extraction, they tend to underperform due to poor coverage of specialized tokens and limited exposure to domain knowledge.

C. Knowledge Distillation for Clinical NLP

Knowledge distillation (KD) [7] has emerged as a powerful technique to compress large models without substantial loss in accuracy. In KD, a smaller student model learns from a larger teacher model by mimicking its probability distributions

or intermediate feature representations. Multi-level distillation frameworks such as TinyBERT [5] and Clinical-TinyBERT extend this concept by aligning both logits and hidden states across multiple layers, enabling deeper knowledge transfer. In the biomedical domain, KD has been successfully applied to create compact yet effective models for resource-constrained environments, such as on-device or real-time clinical applications. However, most existing studies focus on transferring model behavior alone, without considering improvements to token-level representation or vocabulary adaptation, which limits their ability to handle domain-specific terms efficiently.

D. Adaptive Tokenization and Vocabulary Expansion

Tokenization plays a crucial role in how PLMs represent domain knowledge. Standard subword tokenizers (e.g., WordPiece, BPE) are optimized for general text, leading to fragmentation of specialized terms (e.g., "cardiomyopathy" $\rightarrow$ "cardio" + "myo" + "pathy"), which reduces semantic coherence and increases contextual ambiguity. Adaptive tokenization approaches [8], [9] seek to mitigate this by dynamically adding high-frequency domain-specific tokens or retraining the tokenizer on specialized corpora. SciBERT, for example, builds a new vocabulary from scientific text, yielding improved representation for technical terms. Recent works have shown that vocabulary adaptation combined with lightweight fine-tuning can outperform large-scale DAPT in certain specialized domains. However, sequentially combining adaptive tokenization with knowledge distillation for clinical NLP remains underexplored. This gap motivates our approach, which integrates domain-aware vocabulary expansion with a distillation framework to produce a compact yet domain-specialized clinical language model.

III. Methodology

Our approach aims to adapt a lightweight language model, DistilBERT, for efficient and domain-aware clinical text processing. The overall pipeline consists of three main stages: (1) tokenizer expansion using domain-specific token frequency analysis, (2) knowledge distillation from BioClinicalBERT, and (3) fine-tuning on the MACCROBAT dataset for clinical named entity recognition (NER).

A. Tokenizer Extension via Domain Frequency Analysis

To enhance DistilBERT's ability to represent domain-specific medical terminology, we first analyze token distributions in both clinical and general-domain corpora. The clinical corpus comprises biomedical texts such as case reports and PubMed abstracts, whereas the general corpus includes sources like Wikipedia. For each token, we compute its frequency in both corpora and calculate the Kullback–Leibler (KL) divergence to measure the extent to which the token's usage in the clinical domain deviates from general language. Tokens exhibiting high KL divergence are considered characteristic of clinical text and are added to DistilBERT's vocabulary, expanding it from 30,522 to approximately 32,000 tokens. This vocabulary extension reduces the semantic fragmentation of multi-word medical terms (e.g., "cardiomyopathy"), enabling the model to process them as single, coherent units, which in turn improves the performance of downstream medical entity recognition tasks.

B. Knowledge Distillation from BioClinicalBERT

After extending the tokenizer, the next step is to transfer domain-specific knowledge from a high-capacity teacher model, BioClinicalBERT, to the lightweight student model, DistilBERT with the extended vocabulary. This is achieved through *knowledge distillation (KD)*, a process in which the student learns not only from the ground-truth labels but also from the soft predictions of the teacher. The overall loss function is defined as:

$$L_{KD} = \alpha L_{CE}(y, p_s) + (1-\alpha) T^2 L_{CE}(p_t^{(T)}, p_s^{(T)}), \quad (1)$$

where $L_{CE}(y, p_s)$ is the standard cross-entropy loss between the true labels y and the student's predictions p_s, while $L_{CE}(p_t^{(T)}, p_s^{(T)})$ is the cross-entropy between the temperature-scaled teacher and student probabilities. Temperature T smooths the output distributions, highlighting inter-class relationships that are often informative for the student. The weighting factor α balances the contribution of the hard-label loss and the soft-label (teacher) loss.

Intuitively, the student model benefits in two ways: first, it learns to predict the correct labels directly from the training data; second, it mimics the teacher's behavior, capturing richer domain-specific patterns that may not be explicitly encoded in the ground-truth annotations. This dual supervision allows the student to achieve performance closer to BioClinicalBERT while retaining the computational efficiency of DistilBERT. During training, we use the AdamW optimizer with a moderate learning rate, enabling the student to converge effectively while preserving the knowledge imparted by the teacher.

C. Fine-Tuning for Clinical Named Entity Recognition

In the final stage, the distilled model with the extended tokenizer is fine-tuned on the MACCROBAT dataset for clinical named entity recognition (NER). The MACCROBAT corpus contains richly annotated clinical case reports with entity labels such as *Problem*, *Treatment*, and *Test*, making it suitable for evaluating clinical language understanding. We formulate the task as a token-level sequence labeling problem, where each token is assigned a BIO tag (**B**-eginning, **I**-nside, **O**-outside) corresponding to its entity span. The input sequences are tokenized using the updated vocabulary to ensure that domain-specific medical terms (e.g., "hypertension", "myocardial infarction") are represented as single tokens whenever possible.

During fine-tuning, the model is trained using a categorical cross-entropy loss over the token label predictions. We employ the AdamW optimizer with a learning rate of 5×10^{-5} and apply dropout regularization to prevent overfitting. The model is trained for multiple epochs with early stopping based on validation F1 score. This step enables the model to specialize

in recognizing and classifying medical entities within clinical narratives, effectively leveraging both the expanded vocabulary and the distilled knowledge from BioClinicalBERT. The resulting model achieves improved precision and recall for entity categories that are underrepresented or highly domain-specific.

IV. Data Preprocessing

The MACCROBAT dataset consists of clinical case reports annotated with biomedical entities such as *Problem*, *Treatment*, and *Test*. The dataset is distributed in BRAT standoff format, where each document contains a plain-text file (.txt) and a corresponding annotation file (.ann). Each annotation file specifies the entity spans and their associated types but does not include token-level labels. Therefore, a preprocessing pipeline was implemented to convert these span-level annotations into token-level BIO (Begin–Inside–Outside) tags suitable for sequence labeling with transformer models.

A. Conversion from BRAT to BIO Format

Each clinical document was first tokenized using the extended DistilBERT tokenizer. For every annotated entity span in the BRAT file, the corresponding tokens in the text were aligned based on character offsets. Tokens at the beginning of an entity span were labeled with the **B-** prefix, tokens inside the same span were labeled with the **I-** prefix, and all other tokens were assigned the label **O**. This conversion produced a sequence of token-label pairs compatible with standard NER model training. Since the original MACCROBAT dataset only provides entity boundaries and types, this BIO conversion step was crucial for enabling supervised token-level training. During alignment, care was taken to ensure that token boundaries correctly matched entity spans, especially for multi-word medical terms (e.g., 'chronic heart failure"). When subword tokenization occurred, all subwords inherited the label of the original token to maintain consistency. An example of the converted BIO tagging format is shown below:

```
Sentence: The patient was treated
with aspirin for fever.
BIO Tags: O O O O
B-TREATMENT O B-PROBLEM O
```

This conversion pipeline ensured that each token in the dataset was associated with a corresponding label, allowing the model to learn contextual cues for medical entity recognition. By converting the MACCROBAT annotations into token-level BIO format, we enabled direct compatibility with transformer-based NER training frameworks.

B. Text Normalization and Token Alignment

All text was normalized to lowercase to reduce vocabulary sparsity. Punctuation and spacing inconsistencies were standardized, and special characters were retained when medically meaningful (e.g., 'mg", O_2", pH"). Token alignment between the BRAT annotations and the tokenized text was handled using character-level offset matching, ensuring that every annotated span correctly mapped to its corresponding token indices. This process also resolved overlapping entity boundaries where applicable.

C. Dataset Split and Statistics

The preprocessed dataset was divided into 70% training, 15% validation, and 15% test splits, ensuring balanced distribution of entity types across all partitions. The final dataset statistics after preprocessing are summarized in Table I.

TABLE I
MACCROBAT Dataset Overview After Preprocessing

Split	Documents	Entity Types
Training	140	41
Validation	30	41
Test	30	41

V. Experiments and Results

A. Training Setup

Student model fine-tuned with AdamW, learning rate $5e^{-5}$, batch size 16, 5 epochs. KD temperature $T = 2$, $\alpha = 0.5$.

TABLE II
Training Hyperparameters

Hyperparameter	Value
Optimizer	AdamW
Learning Rate	$5e^{-5}$
Batch Size	16
Epochs	5
KD Temperature (T)	2
Loss Weight (α)	0.5

B. Results

TABLE III
Entity Recognition Performance on MACCROBAT Dataset

Model	Precision	Recall	F1	Accuracy
DistilBERT (baseline)	0.81	0.86	0.84	0.92
DistilBERT + Tokenizer + KD	0.85	0.89	0.87	0.95

VI. Discussion

The experimental results demonstrate that combining tokenizer extension with knowledge distillation provides a practical balance between performance and efficiency in clinical NLP. The proposed model achieves an F1 score of 0.87, reducing the performance gap between DistilBERT and BioClinicalBERT while maintaining a significantly smaller size and faster inference time. This indicates that enriching the tokenizer with domain-relevant terms helps reduce semantic fragmentation, allowing the model to capture medical terminology more coherently. Furthermore, knowledge distillation enables the transfer of domain-specific representations from

the teacher model, effectively compensating for the smaller model capacity of the student.

An important observation is that the tokenizer extension alone provided modest improvements, but when combined with knowledge distillation, the gains became more substantial. This synergy suggests that domain-aware lexical coverage and teacher-guided knowledge transfer reinforce each other. Specifically, the extended tokenizer ensures that domain terms are represented as cohesive tokens, while distillation helps the model assign meaningful representations to them. As a result, the model exhibits improved recognition of complex entities such as multi-word symptoms or composite drug names.

Despite these gains, some limitations remain. The approach relies heavily on the quality and representativeness of the domain corpus used for tokenizer extension. Inadequate or biased token frequency estimation could lead to suboptimal vocabulary updates. Additionally, the distillation process depends on the teacher's accuracy and may propagate its biases or overfitting to specific clinical subdomains. Moreover, our evaluation on MACCROBAT, while suitable for proof-of-concept validation, may not fully reflect the diversity of real-world clinical text.

Future research could address these issues by exploring dynamic or adaptive vocabulary learning methods that update token coverage continuously during fine-tuning. Extending this framework to multilingual or cross-institutional settings could also help evaluate its generalizability across different healthcare systems. Finally, incorporating multi-level or self-distillation frameworks could enhance representation alignment between intermediate layers, potentially yielding further performance gains without increasing model size.

VII. Conclusion

We present an efficient, domain-aware DistilBERT model via sequential tokenizer extension and knowledge distillation. Experiments show near BioClinicalBERT accuracy while maintaining efficiency. Future work includes multilingual evaluation and advanced KD strategies.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.

[2] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert: smaller, faster, cheaper and lighter," in *NeurIPS EMC2 Workshop*, 2019.

[3] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," *arXiv preprint arXiv:1904.05342*, 2019.

[4] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: Pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[5] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling bert for natural language understanding," in *EMNLP*, 2020.

[6] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "Mobilebert: a compact task-agnostic bert for resource-limited devices," in *ACL*, 2020.

[7] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[8] V. Sachidananda, A. Suresh, S. Krishnan, and A. Prabhu, "Efficient domain adaptation of language models via adaptive tokenization," *arXiv preprint arXiv:2109.07460*, 2021.

[9] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," in *EMNLP*, 2019.