Efficient Clinical NLP via Adaptive Tokenization and Knowledge Distillation

Advanced Machine Learning - CS4681
Progress Report - M. C. Neluhena (210419F)

1 Introduction

DistilBERT[1] is a lightweight, distilled version of BERT [2], developed using the technique of knowledge distillation[3]. It offers a more compact and efficient architecture while maintaining performance comparable to the original BERT model. However, when applied to specialized domains such as the medical field, DistilBERT's performance drops due to its lack of pretraining on domain-specific corpora. In contrast, domain-adapted BERT variants like ClinicalBERT[4] and BioBERT[5], which are pretrained on large-scale medical texts, demonstrate significantly better performance on tasks such as Medical Entity Recognition (MER). The primary objective of this project is to build a domain-specific, distilled language model by using ClinicalBERT as the teacher model and DistilBERT as the student model. Through knowledge distillation, the aim is to develop a lightweight yet accurate model capable of high performance in medical NLP tasks, particularly Medical Entity Recognition.

2 Literature Review

2.1 Transformer Pretraining for Clinical NLP

Clinical natural language processing (NLP) depends on models that accurately extract medical entities from unstructured text. Transformer-based models such as BERT have achieved state-of-the-art performance but at the expense of high computational cost. DistilBERT offers a smaller, faster alternative but underperforms ClinicalBERT and BioBERT, which are BERT-base models continued on medical corpora.

ClinicalBERT and BioBERT are built on domain-adaptive pretraining (DAPT) using in-domain texts, followed by task-specific fine-tuning. This approach delivers 5–8% higher F1 scores than general pretrained language models (PLMs) on clinical entity recognition tasks[4, 5]. However, DAPT is computationally expensive, often requiring weeks of training on large datasets such as MIMIC-III[6].

2.2 Knowledge Distillation for Model Compression and Domain Transfer

Knowledge distillation (KD), where a smaller "student" model learns from the soft outputs of a larger "teacher," has become a favored approach to compress and adapt models.

- **Clinical Knowledge Distillation (CKD-EHR):** Wang et al.[7] propose distilling Qwen2.5-7B into BERT for clinical disease prediction, achieving 27% F1-score improvement and 22.2x inference speedup on MIMIC-III phenotyping tasks.
- **Adaptive Contrastive KD:** Guo et al.[8] introduce contrastive distillation loss with sample adaptive reweighting for BERT compression, demonstrating effectiveness on GLUE benchmark tasks.
- **Multi-Level KD:** Zhang et al.[9] propose multi-level knowledge distillation for BERT compression, transferring both token-level and relation-level knowledge.

These works establish that KD from domain-pretrained teachers effectively can produce compact models while preserving inference speed.

2.3 Vocabulary and Tokenization Adaptation

General WordPiece tokenizers fragment domain-specific terms, harming semantic meaning. Efficient tokenization strategies mitigate this:

- **Token Distribution Divergence:** Sachidananda et al.[10] measure conditional token distribution divergences between base and domain-specific corpora to efficiently identify domain-specific subword sequences for vocabulary expansion.
- **Vocabulary Expansion:** Sachidananda et al.[10] expand RoBERTa's vocabulary with 10,000 domain-specific tokens, recovering 97% of domain-specific pretraining performance benefits while achieving 72x faster training compared to continued pretraining on domain corpora.

These methods indicate that even small expansions of the vocabulary can lead to significant improvements in domain adaptation.

2.4 Unified Pipelines for Vocabulary Adaptation and Knowledge Distilliation

Despite the independent successes of KD and vocabulary adaptation, few studies jointly optimize both in a single pipeline.

- **Vocabulary-Expansion Knowledge-Distillation:** This integrates vocabulary augmentation with KD, achieving +1% accuracy on biomedical QA and classification benchmarks using 96% less training time than full DAPT(Domain Adaptive Pre Training)[11]. However, its evaluations focus on PubMedQA and BioASQ, not clinical NER.
- **Adapt-and-Distill** Yao et al.[12] proposes a unified pipeline where both teacher and student models are domain-adapted separately before knowledge distillation, achieving the best performance among compression strategies with improvements up to 1.06% F1 over BERT in biomedical tasks, though it requires multi-stage training and separate domain adaptation for both models.

No existing work applies a truly unified, task-specific framework combining adaptive tokenization and KD to clinical NER on datasets like i2b2.

3 Proposed Solution

This project proposes a model that is computationally efficient while aiming to achieve accuracy comparable to or better than ClinicalBERT in Medical Entity Recognition. The expected results will be achieved through the following implementations.

- **Adaptive Tokenization Enhancement:** Extension of DistilBERT's vocabulary with 1,000-2,000 high-frequency clinical tokens, reducing medical term fragmentation and improving semantic representation without expensive domain pretraining.
- **Task-Specific Knowledge Distillation:** Implementation of temperature-scaled knowledge transfer from ClinicalBERT to the enhanced DistilBERT, enabling the compact student model to learn domain expertise from the specialized teacher.

4 Methodology

This section outlines the methodology which will be used to implement the proposed solution.

4.1 Adaptive Tokenization Enhancement

The main goal of this enhancement is to align the model with the clinical domain by capturing the clinical language. The process begins with statistical analysis of token frequency distributions in the clinical corpus compared to general-domain corpora like Wikipedia and BookCorpus. KL divergence is computed for candidate clinical terms to identify tokens that are significantly more frequent in clinical data. Candidate tokens are ranked based on their divergence scores and semantic coherence. Tokens are selected if they meet a minimum frequency threshold , preserve semantic integrity (i.e., represent complete medical terms rather than fragments). Following token selection, the vocabulary of DistilBERT's WordPiece tokenizer is extended from its original size of 30,522 to approximately 32,000 tokens.

4.2 Knowledge Distillation Framework

The core of the model training strategy involves a teacher-student architecture, where ClinicalBERT, fine-tuned on the i2b2 dataset, acts as the teacher model. The student model is a modified version of DistilBERT with an extended clinical vocabulary. The distillation process aims to transfer knowledge from the teacher to the student by learning from soft probability distributions over the NER label space, in addition to the standard supervised learning signals. The training objective is defined as a combination of cross-entropy loss and temperature-scaled knowledge distillation loss.

During training, the DistilBERT model is initialized with the extended vocabulary embeddings, while the ClinicalBERT teacher model is frozen to preserve its learned knowledge. Training is conducted jointly by optimizing both the task-specific loss and the distillation loss. Model convergence is monitored by tracking alignment between hard (ground truth) and soft (teacher-generated) targets.

5 Implementation Timeline

The project will be completed within a five-week period. The weekly timeline is outlined below:

Week 7: Foundation and Baseline Setup

- Acquire and explore the i2b2 2014 dataset and annotation guidelines
- Set up development environment with PyTorch, Transformers, and evaluation tools
- Develop baseline DistilBERT fine-tuning pipeline with preprocessing, tokenization, and evaluation metrics (NER: precision, recall, F1)
- Validate baseline performance on the validation set

Week 8: Core Enhancements (Tokenization & Distillation)

- Develop adaptive tokenization tailored for clinical domain (token distribution analysis, KL-divergence, vocabulary extension)
- Implement and test tokenization improvements on sample clinical texts
- Build knowledge distillation framework (teacher–student training, temperature scaling, combined loss function)
- Integrate ClinicalBERT as teacher model for knowledge transfer

Week 9: Integration and Hyperparameter Optimization

- Integrate adaptive tokenization and knowledge distillation into a unified training pipeline
- Conduct hyperparameter search.

Week 10: Documentation, Validation, and Reporting

- Compile experimental results into performance tables and figures
- Final validation on held-out test set
- Complete code documentation and reproducibility guidelines
- Publishing the completed paper

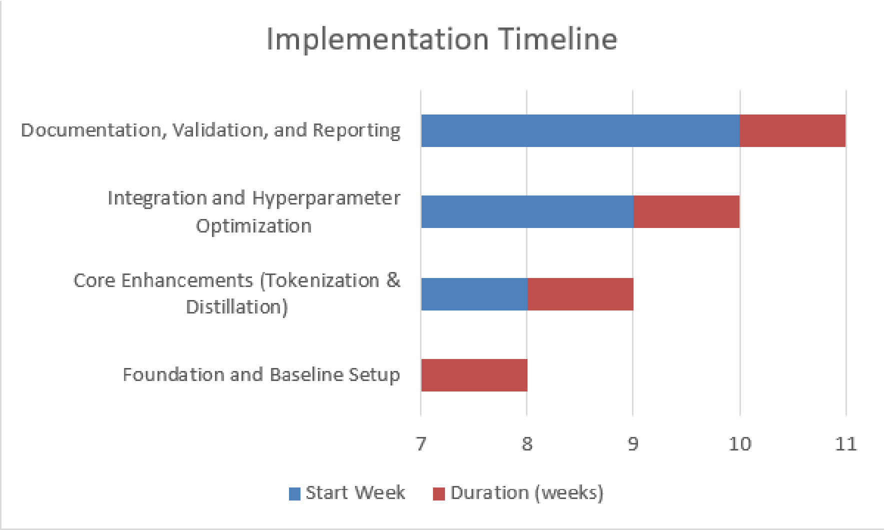

Figure 1: Gantt Chart

References

[1] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," in *NeurIPS EMC2̂ Workshop*, 2019, available: https://arxiv.org/abs/1910.01108.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics, 2019, pp. 4171–4186, available: https://arxiv.org/abs/1810.04805.

[3] C. Buciluă, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2006, pp. 535–541.

[4] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," *arXiv preprint arXiv:1904.05342*, 2019, available: https://arxiv.org/abs/1904.05342.

[5] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020, available: https://academic.oup.com/bioinformatics/article/36/4/1234/5566506.

[6] "Mimic-iii: Medical information mart for intensive care," https://registry.opendata.aws/mimiciii, accessed: 2025-08-24.

[7] J. Wang, H. Ling, L. Zhang, L. Zhang, F. Wang, Y. Gao, and Z. Li, "CKD-EHR: Clinical Knowledge Distillation for Electronic Health Records," *arXiv preprint*, vol. arXiv:2506.15118, Jun. 2025, arXiv, 2025-06-18.

[8] J. Guo, J. Liu, Z. Wang, Y. Ma, R. Gong, K. Xu, and X. Liu, "Adaptive contrastive knowledge distillation for BERT compression," in *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 8941–8953.

[9] Y. Zhang, Z. Yang, and S. Ji, "Mlkd-bert: Multi-level knowledge distillation for pre-trained language models," 2024. [Online]. Available: https://arxiv.org/abs/2407.02775

[10] V. Sachidananda, J. S. Kessler, and Y. Lai, “Efficient domain adaptation of language models via adaptive tokenization,” *CoRR*, vol. abs/2109.07460, 2021. [Online]. Available: https://arxiv.org/abs/2109.07460

[11] P. Gao, T. Yamasaki, and K. Imoto, “VE-KD: Vocabulary-expansion knowledge-distillation for training smaller domain-specific language models,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 15 046–15 059. [Online]. Available: https://aclanthology.org/2024.findings-emnlp.884/

[12] Y. Yao, S. Huang, W. Wang, L. Dong, and F. Wei, “Adapt-and-distill: Developing small, fast and effective pretrained language models for domains,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 460–470. [Online]. Available: https://aclanthology.org/2021.findings-acl.40/