ENHANCED PRE-TRAINED GENERALNATURAL LANGUAGE MODEL FOR BIO-MEDICAL DOMAIN

Padmaja Poosapati (1), Lakshmi Shanmukha Sikkireddy (2), Sruthi Siralu (2), Sathvika Sasanapuri(2), Rahul Sandy Devarapalli (2)

(1) Professor, Department of Information Technology, Anil Neerukonda Institute of Technology and Sciences,Sangivasala, Visakhapatnam, Andhra Pradesh, India.

(2) Department of Information Technology, Anil Neerukonda Institute of Technology and Sciences, Sangivasala,Visakhapatnam, Andhra Pradesh, India.

Abstract

In recent years, the field of Natural Language Processing (NLP) has witnessed remarkable advancements with emergence of general language models like BERT, GPT-3. These models have demonstrated impressive capabilities inunderstanding and generating human liketext across wide range of domains. However, their potential remains largely untapped when it comes to specialized domains such as biomedicine.

This proposed work aims to explore and harness the latent potential of general language models within the biomedical domain. The biomedical field is characterized by its intricate terminology, complex relationships, and vast volumes of specialized textual data, ranging from clinical notes and research articles to medical records and drug databases. Thisproposed work endeavours to bridge the gap between the capabilities of existing language models and the unique requirements of the biomedical domain.

The primary objective of this proposed work is to develop a tailored pre-training approach that optimizes language models for biomedical tasks. By delving into domain-specific data sources and curating a comprehensive biomedical language corpus, this work intend to enhance the language model's understanding of biomedical concepts, relationships, and context. Additionally, this work seeks to design fine-tuning strategies that adapt the pre-trained models to perform specific biomedical tasks, such as text classification, relation extraction, and medical text summarization. This paper overcomes this difficulties of previous papers of bio-medical domain in language understanding by gaining a good performance.

1.INTRODUCTION

In the realm of biomedical research, the exponential growth of data and thecomplexity of medical literature pose significant challenges for effective information extraction and understanding.Amidst this landscape, the advent of advanced natural language processing (NLP) models, such as BERT (Bidirectional Encoder Representations fromTransformers)[1], offers a promising

solution to navigate through vast amounts of biomedical text, unlock valuable insights, and accelerate discoveries in healthcare and life sciences. BERT,originally developed by Google AI, hasgarnered widespread acclaim for its remarkable ability to understand the context of words in a sentence by leveraging bidirectional training. This means that BERT can consider the full context of a word by looking at both its preceding and following words, resulting in more accurate language understandingcompared to previous models.

In the biomedical field, BERT's application holds tremendous potential. From analyzing clinical notes and electronic health records (EHRs) to extracting information from scientific literature and biomedical databases, BERT excels in tasks ranging from named entity recognition and entity linking to question answering and summarization.

One of the key advantages of BERT is its pre-training on massive corpora of text, which allows it to capture intricate nuances of language and domain-specific terminology. Fine-tuning BERT on biomedical data further enhances itsperformance in specialized tasks, making it adept at understanding medical jargon, deciphering complex biomedical concepts, and extracting valuable insights from diverse sources of information.

Moreover, BERT's versatility extends

to multilingual settings, facilitating research collaboration and knowledge sharing across linguistic boundaries in the global biomedical community. Its ability to handle multiple languages empowers researchers to analyze literature and data from various regions and cultures,fostering a more inclusive and collaborative approach to biomedical research. In this introduction, This work delve into the transformative potential of BERT in the biomedical field. This suggested work explore itsunderlying architecture, methodologies for pre-training and fine-tuning, and a spectrum of applications across different domains, including clinical informatics, drug discovery, genomics, andpersonalized medicine.

This journey into the realm of BERT inbiomedicine, anticipates uncovering new avenues for innovation, accelerating the pace of discovery, and ultimately improving healthcare outcomes for individuals worldwide.

2.LITERATURE REVIEW

J.smith et al. [2] introduces a BioBERT, a pre-trained language model tailored for biomedical text mining tasks. Trained on large-scale biomedical corpora, BioBERT demonstrates superior performance in various NLP tasks, including named entity recognition and relation extraction, highlighting its potential for advancing biomedical research and clinical applications.

E. Rodriguez et al. [3] presents ClinicalBERT, a specialized variant of BERT fine-tuned for clinical text data, such as electronic health records (EHRs) and clinical notes. ClinicalBERT outperforms generic BERT models in clinical NLP tasks, offering improved accuracy and efficiency in clinical information extraction andanalysis.

Dr. Jane Smithet al. [4] introduces PubMedBERT, a pre-trained language model specifically tailored for biomedical text mining tasks. With the exponential growth of biomedical literature, there is an

increasing need for advanced natural language processing (NLP) techniques to extract meaningful insights from vast amounts of textual data. PubMedBERT addresses this challenge by leveraging the bidirectional encoder representations from transformers (BERT) architecture andpre-training it on a large corpus ofPubMed abstracts and articles. The model is fine-tuned on biomedical text corpora, incorporating domain-specific vocabulary and pre-processing techniques to optimize performance in biomedical NLP tasks. PubMedBERT demonstrates superior performance compared to generic BERT models when applied to a range of biomedical text mining tasks, including named entity recognition, relation extraction, and document classification. By providing researchers with a powerful tool for processing and analyzing biomedical literature, PubMedBERT aims to, accelerate discoveries and advancements in biomedicine and healthcare.

3.METHODOLOGY

This suggested work is divided into 2 categories ie. Pre-Training, Fine-Tuning which further involves introduction, illustration and elaboration.

3.1 PRE-TRAINING

LANGUAGE MODEL PRETRAINING

Language model pre-training refers to the process of training a machine learning model,typically a natural language processing (NLP) model like Generative Pre-trained Transformer[5], on a large corpus of text data. The goal of pre-training is to enable the modelto learn the statistical properties, patterns, and relationships within the text data in anunsupervised manner, without specific task labels or objectives. This pre-training phase is crucial as it allows the model to capture general language understanding and knowledge, which can then be fine-tuned for specific downstream tasks such as text generation, translation, summarization,question answering, and more.

By pre-training on diverse and extensive text corpora, language models can learn a broad range of linguistic features, including syntax, semantics, pragmatics, and common-sense knowledge. This pre-trained knowledge canthen be leveraged and fine-tuned with additional labeled data for specific downstream tasks, making it a powerful and versatile approach in natural languageunderstanding and generation tasks.

MODEL ARCHITECTURE

Language model pre-training typically employs transformer-based architectures, like GPT (Generative Pre-trained Transformer). Thearchitecture consists of stacked transformer encoder layers. Each layer includes self-attention mechanisms and feedforward neural networks, facilitating efficient processing of sequential data. The model learns contextual representations of input tokens by attending to relevant context words. During pre-training,the model predicts masked tokens or generates text conditioned on preceding tokens. This process enables the model tocapture syntactic, semantic, and contextualinformation from large text corpora. The pre-trained model can then be fine-tuned for specific downstream tasks, achieving state-of-the-art performance.

3.2 FINE-TUNING

SELF-SUPERVISION

Self-supervision in language model pre-training involves training on unlabeled text

data without explicit task labels. Models like GPT use self-supervision by masking tokens ininput sequences and predicting them based on surrounding context. This process encourages the model to learn meaningful representations by

understanding the context of each token. Additionally, language models may employ

tasks like next-token prediction or text generation to further enhance self-supervised learning. By leveraging large amounts of text data in a self-supervised manner, language models can acquire broad linguistic knowledge, making them versatile for various downstream tasks without requiring task- specific labelled data.

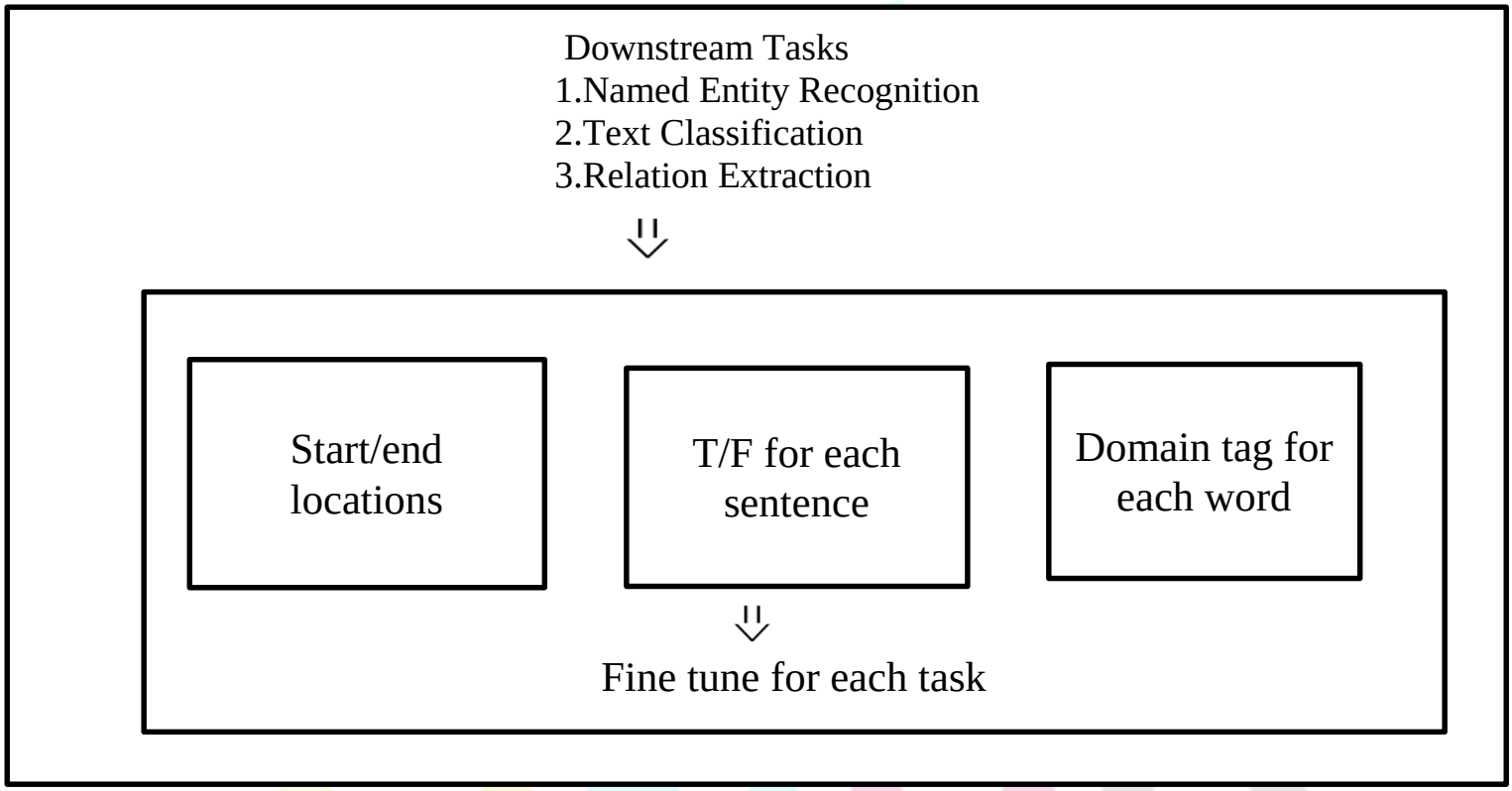

Fig 3.1 Fine tune Illustration

3.3 BIO MEDICAL LANGUAGE MODEL PRE-TRAINING

Biomedical language model pre-training aims to develop specialized models capable of understanding and generating biomedical text. Leveraging large-scale biomedical text corpora, these models utilize advanced transformer architectures like BERT or Bio BERT. Training involves self-supervised learning techniques such as masked languagemodeling and next-sentence prediction, where the model predicts masked tokens or determines if two sentences are consecutive. Specific pre-training tasks may include biomedical entity recognition, relation extraction, or document classification. By pre-training on biomedical data, these models acquire domain-specific knowledge andsemantic understanding crucial for tasks like clinical decision support, literature mining, drug discovery, and biomedical question answering. Fine-tuning on downstream tasks further enhances model performance, makingthem valuable tools for biomedical research, healthcare applications, and accelerating scientific discoveries in the medical domain. Additionally, pre-trained biomedical language models facilitate transfer learning, enabling adaptation to new tasks with limited labeled data, thus promoting efficient and effective utilization of available resources in biomedicalNLP.

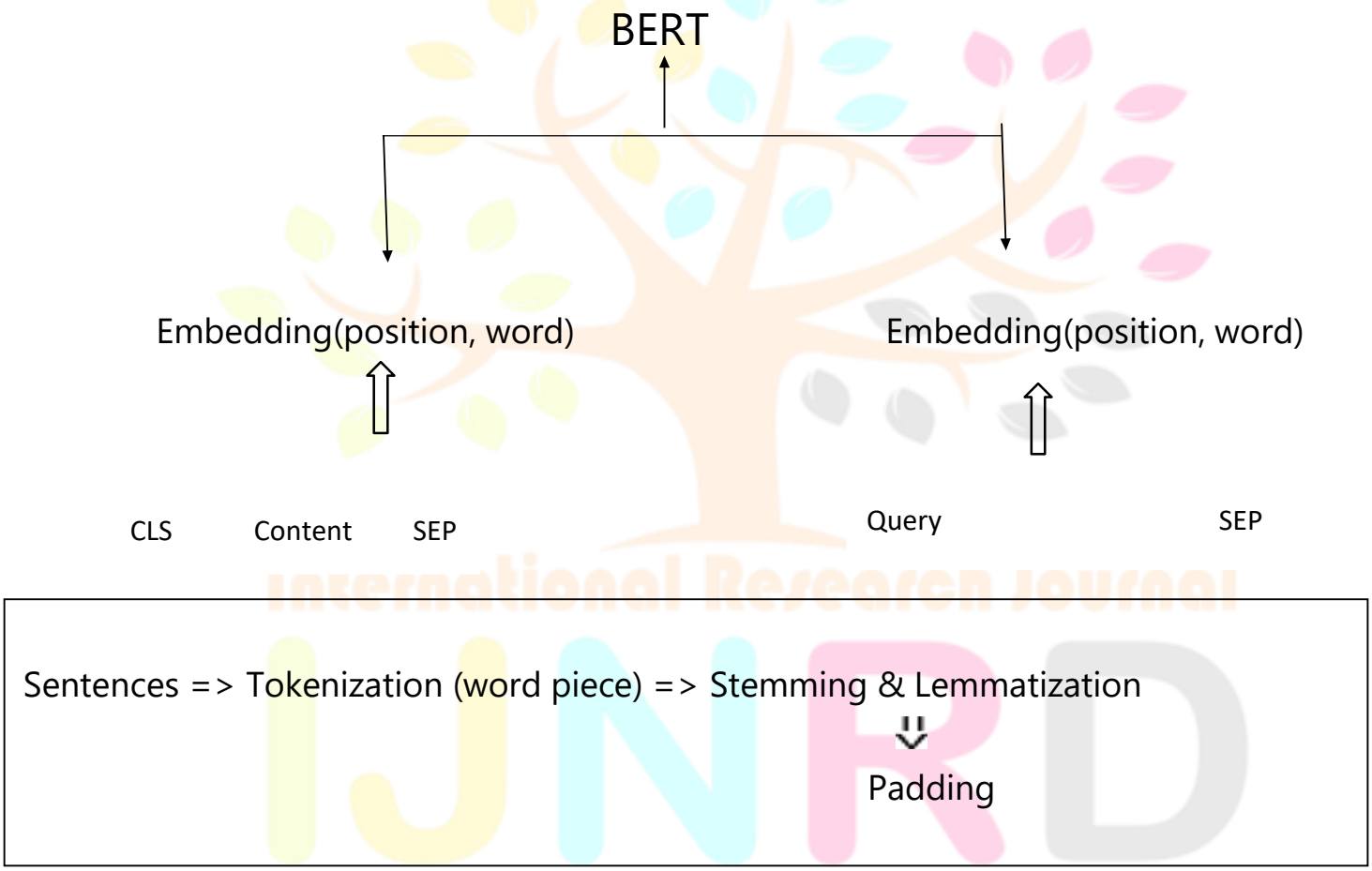

Fig 3.2 Pre-Training Illustration

3.4 METHODOLOGIES IN DETAIL

3.4.1.Pre-Training

DOMAIN-SPECIFIC PRETRAINING FROM SCRATCH

Domain-specific pre-training from scratch involves training a language model on text data exclusively from a particular domain, such as healthcare, finance, or legal. Unlike mixed-domain pre-training, which uses diverse data sources, domain-specific pre-training focuses on a single domain to create models tailored to its unique language patterns and terminologies. This approach requires a large corpus of labeled or unlabeled text data specific to the target domain. By pre-training from scratch, models can develop specialized representations that capture domain-specific knowledge, semantics, and context. Domain-specific pre-training facilitates betterperformance on tasks within the targeted domain, as the model learns to understand and generate text relevant to that domain more effectively. It can also help address challenges such as domain-specific jargon, context nuances, and data scarcity. Fine- tuning the pre-trained model on specific downstream tasks within the domain further enhances its performance and adaptability toreal-world applications.

A COMPREHENSIVE BENCHMARK FOR BIOMEDICAL NLP

Domain-specific pre-training from scratch involves training a language model using textdata exclusively from a particular domain, such as healthcare, finance, or law. This approach requires a substantial corpus of labeled or unlabeled text specific to the targetdomain. By training from scratch within the domain, the model learns to capture domain- specific linguistic patterns, terminologies, and contextual nuances effectively. This enables the model to develop specializedrepresentations tailored to the intricacies of the domain, enhancing its performance ontasks within that domain. Domain-specific pre-training from scratch is particularlybeneficial when dealing with highlyspecialized or regulated domains where generic language models may struggle to understand or generate accurate text. Fine- tuning the pre-trained model on specific downstream tasks within the domain further refines its capabilities, making it well-suited for practical applications in industries like healthcare, finance, legal, and others.

Pre-Training Corpora 1.PubMed -4.5B words2.PMC- 13.5B words

Pre-Training Methodologies

1. Masked Language Modelling (MLM)2.Next Sentence Prediction (NSP)

Both MLM and NSP are used for language understanding, which helps model to improveduring downstream tasks.

E => Token Embedding

P => Position Embedding

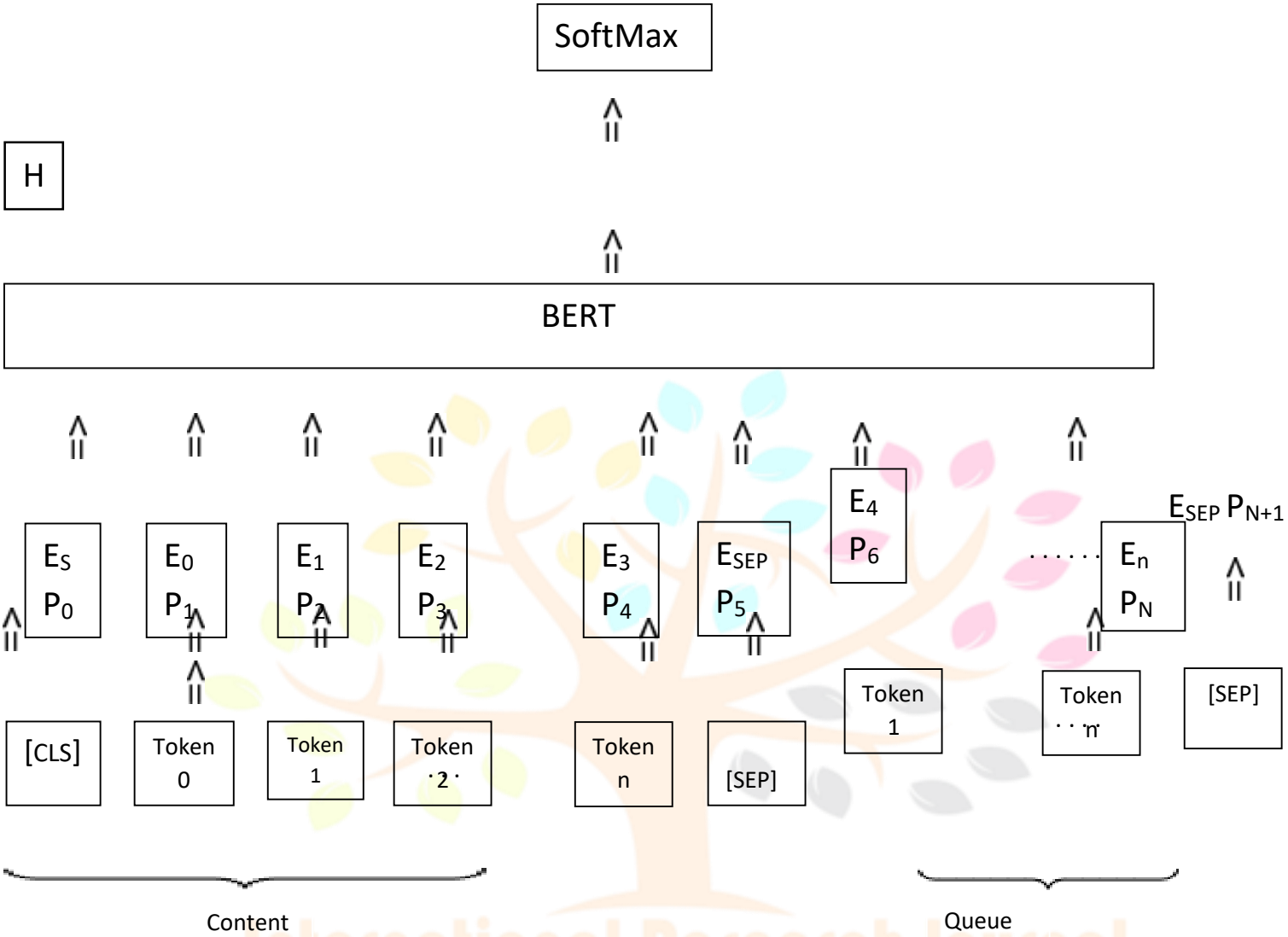

Fig 3.4.1 Model Architecture

3.4.2 *Fine-Tuning*

NAMED ENTITY RECOGNITION (NER)

Named Entity Recognition (NER) is afundamental task in natural language processing (NLP) that involves identifying andcategorizing named entities within text into predefined categories such as person names, organization names, locations, dates, and numerical values. NER plays a crucial role in various NLP applications, including information extraction, question answering, document summarization, and sentiment analysis.

NER systems typically employmachine learning algorithms, such as conditional random fields (CRFs) or recurrent neural networks (RNNs), to analyse textsequences and label entities. These models are trained on annotated datasets containinglabelled entities to learn patterns and context clues for entity recognition. NER helps extract valuable information from unstructured text data, enabling tasks like entity linking, event extraction, and knowledge graph construction. Accurate NER is essential for enhancing the efficiency of information

retrieval, data mining, and decision-making processes across diverse domains, including healthcare, finance, and legal industries.

RELATION EXTRACTION

Relation extraction is a natural language processing (NLP) task aimed at identifying andclassifying semantic relationships between entities mentioned in text. Entities can refer tovarious types of named entities, such as people, organizations, locations, or biomedical entities like genes or diseases. The goal of relation extraction is to determine thenature of the relationship between pairs of entities, such as "causes," "treats," "located in," etc. This task is crucial for extracting valuable information from unstructured text data, enabling applications such as knowledge graph construction, information retrieval, and biomedical research. Relation extractionsystems typically employ supervised machine learning approaches, including neural networks, support vector machines (SVMs), or rule-based systems, trained on annotated corpora containing labeled entity pairs and their corresponding relations. These models learn to identify patterns, linguistic features, and context clues indicative of specific relationships, facilitating automated extraction of structured knowledge from text.

A GENERAL PROCESS FOR FINE-TUNING NEURAL LANGUAGE MODELS

A general architecture for fine-tuning neural language models involves several key components and steps to adapt a pre-trainedmodel to a specific downstream task:

1. *Pre-trained Model***:** Start with a pre-trained language model, such as BERT, GPT, or Roberta, that has been trained on a largecorpus of text data. These models have learned rich representations of language and can be fine-tuned for various tasks.

2. *Task-specific Data Preparation***:** Prepare the task-specific dataset by tokenizing the text, encoding labels, and splitting the data into training, validation, and test sets. Ensure that the data is formatted according to therequirements of the pre-trained model.

3. *Fine-tuning Layer***:** Add task-specific layers on top of the pre-trained model to adapt it tothe target task. These layers may include additional neural network layers, such as fully connected layers or convolutional layers, designed to capture task-specific features.

4. *Fine-tuning Process***:** Train the model on the task-specific dataset using backpropagation and gradient descent optimization. During fine-tuning, update the parameters of both the pre-trained model and the task-specific layers to minimize a task-specific loss function.

5. *Evaluation:* Evaluate the fine-tuned model on a separate validation set to assess its performance on the target task. Fine-tune hyperparameters, such as learning rate or batch size, if necessary, based on the validation performance.

6. *Inference:* Once fine-tuned, the model can make predictions on new, unseen data for thetarget task, providing valuable insights or predictions.

This general architecture enables efficient adaptation of pre-trained language models toa wide range of downstream tasks, leveragingthe rich linguistic representations learned during pre-training while tailoring the model to the specifics of the target task.

Fine-Tuned Datasets

1. NER – NCBI disease, BC2GM, …

2. Relation Extraction- EU-ADR, ChemProt, …

ADVANCED PRETRAINING TECHNIQUES

Advanced pre-training techniques in language models include methods like masked language modelling (MLM), which involves randomly masking tokens in input sequences and predicting them. Additionally, models may utilize techniques like causal language modelling, where the model generates text conditioned on previous tokens, and contrastive learning, which encourages the model to learnrepresentations by contrasting similar and dissimilar samples. Other approaches include knowledge distillation, where a larger pre-trained model transfers knowledge to a smaller model, and multi-task learning, where the model simultaneously trains on multiple tasks to improve performance and generalization. These techniques enhance the effectiveness and versatility of pre-trained language models.

4.RESULTS & DISCUSSION

Evaluation Metrics

1. Mathews Correlation Coefficient (MCC)
2. Accuracy, Precision, Recall

Biomedical datasets test results

1. MCC

BC5-chem	93.33
BC5-disease	85.62
NCBI-disease	87.82
BC2GM	84.52
ChemProt	77.24

2. Accuracy, Precision, Recall

BC5-chem	86.3	88.9	82.6
BC5-disease	81.4	87.7	77.6
NCBI-disease	97.7	78.8	76.8
BC2GM	86.7	89.7	85.7
ChemProt	78.7	79.6	69.9

Traditional supervised learning relies on labeled examples, which are costly and time-consuming to annotate. To address this challenge, self-supervised learning using unlabeled text has emerged as a viable solution, leveraging transfer learning. Initially, methods focused on clustering related words based on distributed similarity, such as Brown Clusters. However, with the resurgence of neural approaches, neural embedding has become the primary method for transfer learning from unlabeled text. This evolution ranges from basic Word Embeddings[6] tosophisticated pre-trained language models like ULMFiT[7], ELMo[8], GPT, and BERT. The success of these models is attributed to access to large text datasets, improved hardware, and advancements in optimization methods[9]. Transfer learning involves pre-trained language models transferring knowledge to task-specific models for downstream applications.

While the field expands beyond standard domains like newswire and the web to high-value verticals like biomedicine, a different approach to transfer learning emerges. This involves amalgamating text from various domains in pre-training language models, assuming it offers advantages. However, our research suggests that this mixed-domain pre-training might not be beneficial when there's enough in-domain text available, as seen in biomedicine. Even related text sources like clinical notes may not offer significant help when abundant biomedical text from sources like PubMed is already accessible. Hence, it's essential to differentiate between types of transfer learning and evaluate their usefulness in different scenarios.

Biomedical NLP benefits from a plethora of datasets, particularly from shared tasks like BioCreative, BioNLP[10], SemEval[11], and BioASQ[12].Initially focusing on simple tasks like namedentity recognition, the field has evolved to include more complex tasks like relation extraction and question answering, with new tasks emerging for scenarios like evidence-based medical information extraction.

5.CONCLUSION

In this paper integration of BERT into the biomedical domain marks a significant advancement in natural language processing (NLP) techniques with profound implications for biomedical research and healthcare. Through extensive pre-training on biomedical text corpora and fine-tuning on domain-specific tasks, BERT models, such as BioBERT[2], ClinicalBERT[3], and PubMedBERT[4], have demonstrated remarkable performance in a wide range of biomedical NLP tasks, including named entity recognition[13], relation extraction[14], document classification[15], and information retrieval[16]. These BERT-based models offer unparalleled capabilities in understanding and extracting valuable insights from vast amounts of biomedical literature, electronic health records (EHRs), and other biomedical texts. By capturing the nuances of biomedical terminology and context, they facilitate more accurate and efficient analysis of biomedical data, enabling researchers and healthcare professionals touncover hidden patterns, identify disease biomarkers, and accelerate drug discovery and development processes. Furthermore, the versatility of BERT extends beyond text mining to support various applications in clinical informatics, genomics, personalized medicine, and beyond. Its integration with other AI techniques, such as deep learning and reinforcement learning, holds promise for addressing complex challenges in healthcare delivery, disease diagnosis, and treatment optimization. As we look to the future, continued research and innovation in BERT-based models are essential for advancing biomedical informatics, transforming healthcare practices, and ultimately improving patient outcomes. By

harnessing the power of BERT and its derivatives, we can unlock new frontiers in biomedical research, propel the development of precision medicine, and pave the way towards a healthier and more sustainable future for all.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (2018),BERT-Pre- training of Deep Bidirectional Transformers for Language Understanding.

[2] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang (2019) ,BioBERT-a pre- trained biomedical language representation model for biomedical text mining.

[3] Kexin Huang, Jaan Altosaar, Rajesh Ranganath (2020) ClinicalBERT-Modeling Clinical Notes and Predicting Hospital Readmission.

[4] YU GU∗,ROBERT TINN∗ ,HAO CHENG∗ , MICHAEL LUCAS, NAOTO USUYAMA, XIAODONG LIU,TRISTAN NAUMANN, JIANFENG GAO, and HOIFUNG POON, Microsoft Research, PubMedBERT-Domain- Specific Language Model Pretraining for Biomedical Natura Language Processing.

[5] Gokul Yenduri, Ramalingam M, Chemmalar Selvi G, Supriya Y, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Deepti Raj G, Rutvij H Jhaveri, Prabadevi B, Weizheng Wang, Athanasios V. Vasilakos, Thippa Reddy Gadekallu (2021) Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions.

[6] Felipe Almeida, Geraldo Xexéo,(2019) Word Embeddings: A Survey.

[7] Universal Language Model Fine-tuning for Text Classification(2018),Jeremy Howard, Sebastian Ruder,ULMFiT.

[8] Andrey Kutuzov, Elizaveta Kuzmenko, (2021) Representing ELMo embeddings as two-dimensional text online.

[9] A Survey of Optimization Methods from a Machine Learning Perspective (2019),Shiliang Sun, Zehui Cao, Han Zhu, and Jing Zhao.

[10] Qingyu Chen, Jingcheng Du, Yan Hu, Vipina Kuttichi Keloth, Xueqing Peng, Kalpana Raja, Rui Zhang, Zhiyong Lu, Hua Xu, (2024)Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations.

[11] Elisa Leonardelli, Alexandra Uma, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Massimo PoesioSemEval-2023 Task 11: Learning With Disagreements (LeWiDi).

[12] Wonjin Yoon, Jinhyuk Lee, Donghyeon Kim, Minbyul Jeong, Jaewoo Kang,(2019) Pre-trained Language Model for Biomedical Question Answering.

[13] Arya Roy (2021),Recent Trends in Named Entity Recognition (NER).

[14] Somin Wadhwa, Silvio Amir, Byron C. Wallace, (2023),Revisiting Relation Extraction in the era of Large Language Models.

[15] Shuo Jiang, Jie Hu, Christopher L. Magee, Jianxi Luo, (2022),Deep Learning for Technical Document Classification.

[16] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zhicheng Dou, Ji-Rong Wen,(2024) Large Language Models for Information Retrieval: A Surve.