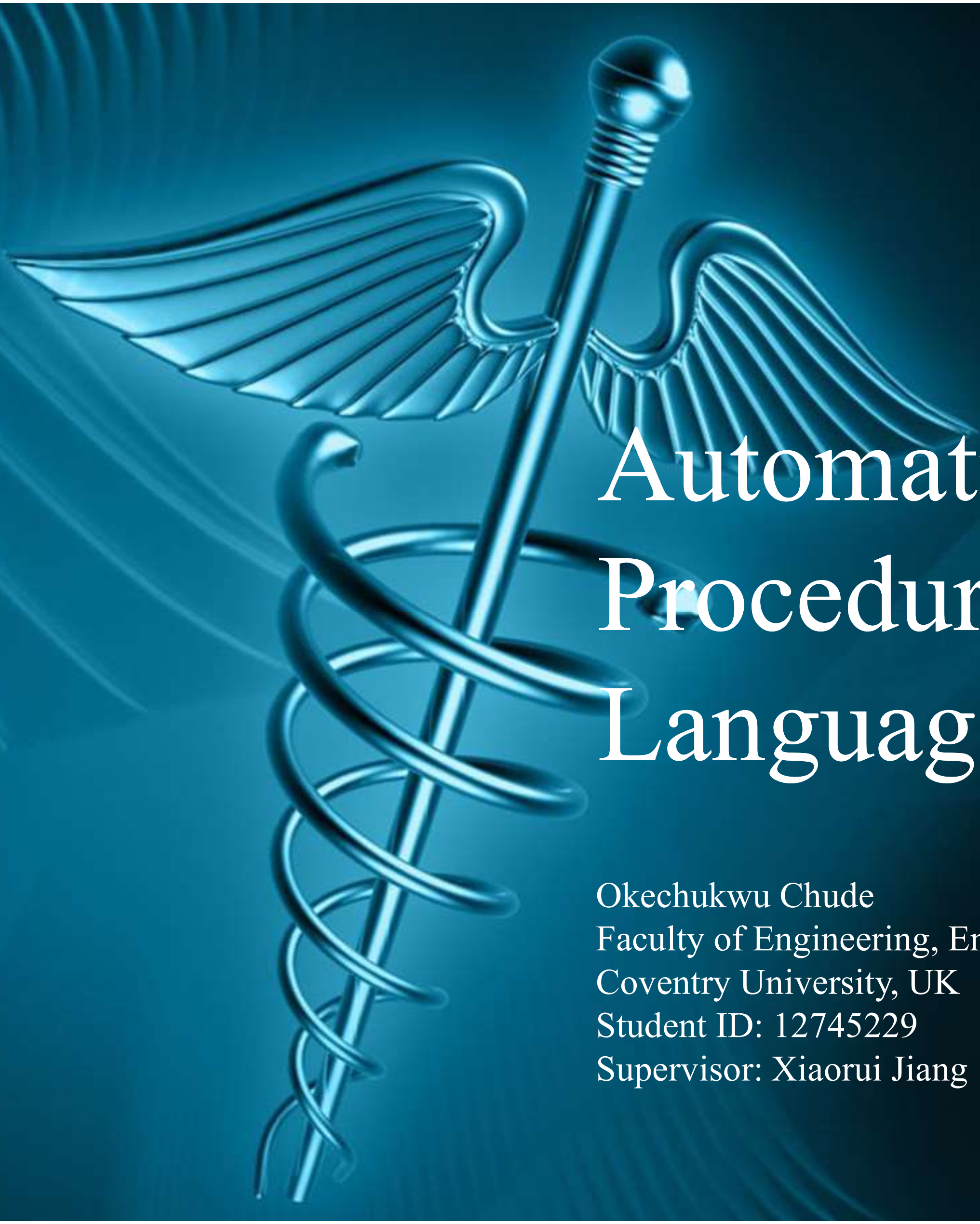
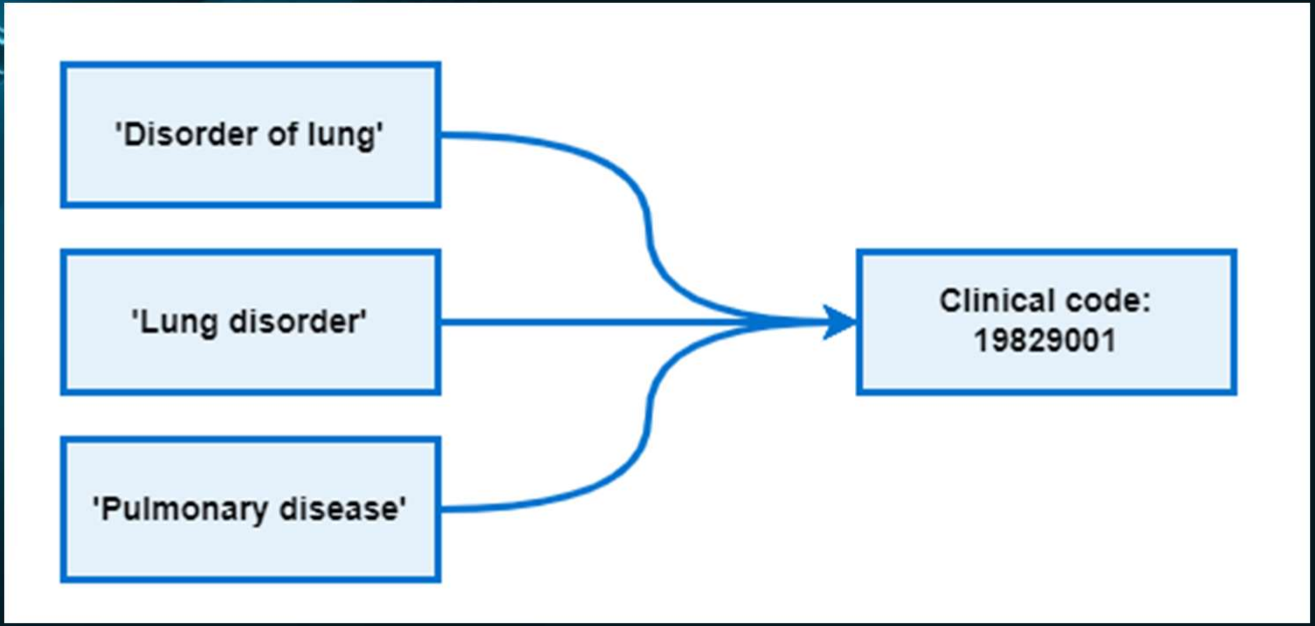


Automating Diagnosis and Procedure Coding with Natural Language Processing

Okechukwu Chude
Faculty of Engineering, Environment and Computing
Coventry University, UK
Student ID: 12745229
Supervisor: Xiaorui Jiang

Introduction- What is Diagnosis and Procedure Coding?

- In the medical field, Diagnosis and Procedure Coding (Medical coding) is the process of translating various aspects of healthcare—such as diagnoses, procedures, medical services, and equipment—into universal medical alphanumeric codes.



'Disorder of lung'

'Lung disorder'

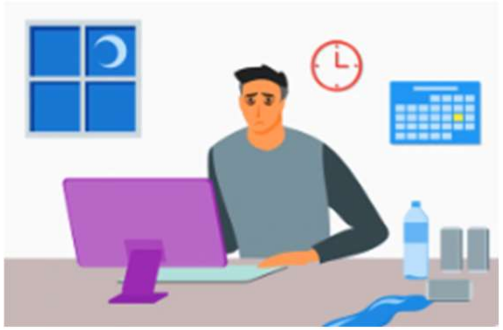
'Pulmonary disease'

Clinical code: 19829001

Fig.1 Clinical Coding structure

Challenges faced in clinical coding

Error rates for primary discharge diagnoses and principal procedure coding in clinical coding are well documented to range from 8% to over 31% [2]



To satisfy billing expectations, coders frequently labour under pressure. Inaccuracies and omissions may occur due to the need to code rapidly, which could lower reimbursement rates.

Shifting Responsibilities: Because medical coding is done by non-clinical office workers, the documentation from the doctor is more important in the abstracting of data process.

Electronic health records can help with documentation, but they also bring with them problems including inconsistent data entry, a lack of training and upskilling, and incompatibility with other data.

Problem Statement

- Medical coding conducted manually is a laborious and error-prone operation.
- To assign the correct diagnostic and procedure codes from unstructured clinical records, skilled human coders are needed.
- This procedure is costly, time-consuming, and prone to errors and inconsistencies because of coding system variability and human error.
- Its inability to scale makes it difficult to code massive amounts of data for population health management and research.

Research Questions

- What NLP techniques (e.g., named entity recognition, relation extraction) are most effective for extracting relevant evidence from clinical narratives for coding purposes?
- How effective are current NLP techniques in accurately extracting diagnostic and procedural information from unstructured clinical text?
- What are the best practices for evaluating the performance and accuracy of NLP systems designed for automating diagnosis and procedure coding tasks in healthcare?

Aims & Objectives

The goal of this project is to automate the process of assigning procedure and diagnosis codes to medical records by utilizing Natural Language Processing (NLP) techniques.

This project will produce a model that automates the process of assigning diagnosis and procedure codes in the healthcare industry, resulting in major gains in productivity, precision, and data quality.

Objectives

- Data Collection and Preprocessing – Obtain existing labelled data for training and testing and apply the following preprocessing techniques; including tokenization, cleaning, normalization, etc.
- Developing the Model – Develop a model for extracting medical terms from clinical text and mapping to their associated code.
- Performance evaluation – Comparing the model's performance to the currently existing model’s using metrics like precision, recall, and F1-score to quantify the model's ability to correctly identify and assign relevant codes.

Related Works

- Cheng et al. (2023) introduces MDACE, the first publicly available code evidence dataset built on a subset of the MIMIC-III clinical records, which contains evidence spans for diagnosis and procedure codes annotated by professional medical coders.
- **Strengths**
  - Provides evidence annotations for long documents in an extreme multi-label classification setting
  - Offers a dataset that can be used to evaluate and enhance the explainability of deep learning models
- **Weaknesses**
  - Limitations in the annotation process, such as potential bias in finding sufficient evidence and inconsistencies in cleaning up the data.
  - Limited overlap between the MIMIC codes and MDACE codes, indicating missing codes in the annotations
  - Potential impact of input text truncation on evidence results

- Huang et al. (2022) addresses the challenge of automatically classifying electronic health records (EHRs) into diagnostic codes using pretrained language models (PLMs).
- **Strengths**
  - The paper identifies and analyses the main challenges of applying PLMs to automatic ICD coding, including long text input, large label sets, and domain mismatch
  - The authors develop various techniques within the PLM-ICD framework to tackle these challenges and overcome the underperformance of pretrained language models on this task
- **Weaknesses**
  - The paper does not provide a comparison with other state-of-the-art methods in the field of automatic ICD coding, which limits the ability to fully assess the superiority of the proposed PLM-ICD framework

- The paper does not thoroughly discuss the limitations or potential drawbacks of using pretrained language models for ICD coding, which could provide a more comprehensive understanding of the approach
- The paper focuses primarily on the MIMIC dataset, and it would be beneficial to evaluate the proposed framework on other datasets to assess its generalizability and robustness

Methodology (BERT)

Bidirectional Encoder Representations from Transformers (BERT) is a language model based on the transformer architecture. It is a research published by researchers at Google AI Language in 2018.

- BERT’s key technical innovation is applying the bidirectional training of Transformer, a popular attention model, to language modelling.

- This contrasts with previous efforts which looked at a text sequence either from left to right or combined left-to-right and right-to-left training.

- The paper’s results show that a language model that is bi-directionally trained can have a deeper sense of language context and flow than single-direction language models.

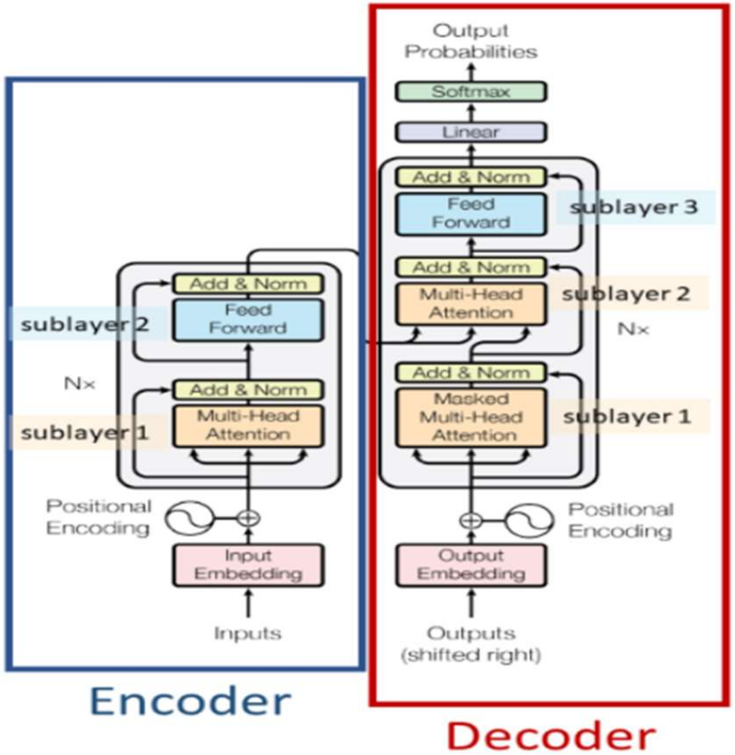


Fig2. Transformer Architecture

- BERT eliminates the Decoder component of the Transformer architecture and solely utilises the Encoder section.

- Its encoder architecture is identical to that of the Transformer. BERT, to put it briefly, is essentially the stacking of encoders

Methodology (BioBERT)

- BioBERT is a contextualized language model, based on BERT.

- The paper is from researchers of Korea University & Clova AI research group based in Korea.

- The significant contribution is a pre-prepared bio-clinical language representation model for different bio-medical text mining tasks.

Development Process

- uses a pre-trained BERT model that was trained on Wikipedia to initialise BioBERT. Books Corpus has 0.8 billion words and 2.5 billion words.

- Next step was pre-training on the domain data; in this case, BioBERT was pre-trained on 4.5 billion words from PubMed Abstracts and 13.5 billion words from PMC Full-text articles.

- NER, Q&A, and connection extraction are just a few of the biomedical text mining tasks that the pre-trained model was used to hone on.

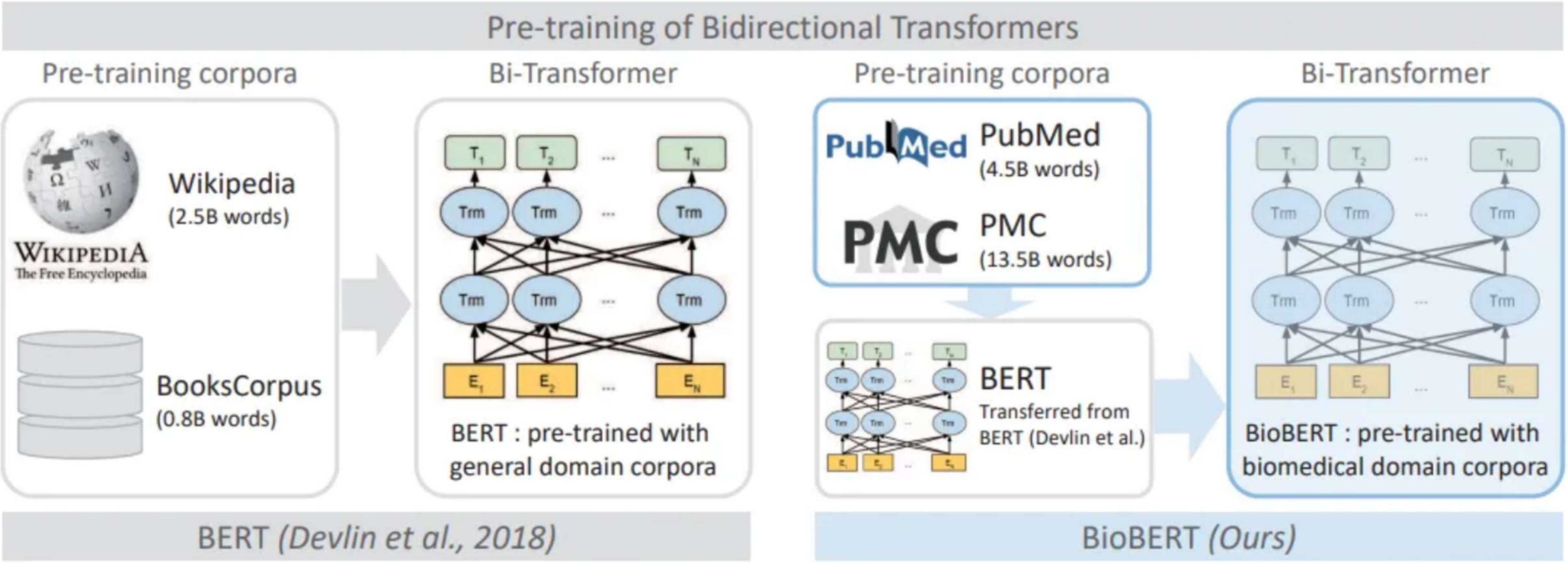


Fig 3 BioBERT development

Results

BERT Performance

Label	Precision	Recall	F1-Score
B	0.73	0.76	0.74
I	0.75	0.7	0.73
O	0.97	0.97	0.97

BioBERT Performance

Label	Precision	Recall	F1-Score
B	0.74	0.71	0.73
I	0.75	0.69	0.72
O	0.96	0.97	0.97

References

1. *NHS choices*. Available at: https://digital.nhs.uk/developer/guides-and-documentation/building-healthcare-software/clinical-coding-classifications-and-terminology (Accessed: 21 March 2024).
2. Naran, S., Hudovsky, A., Antscherl, J., Howells, S., & Nouraei, S. A. R. (2014). Audit of accuracy of clinical coding in oral surgery. *British Journal of Oral and Maxillofacial Surgery*, *52*(8), 735–739. https://doi.org/10.1016/j.bjoms.2014.01.026
3. Cheng, H. *et al.* (2023) 'MDACE: Mimic documents annotated with code evidence', *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* [Preprint]. doi:10.18653/v1/2023.acl-long.416.

4. Huang, C.-W., Tsai, S.-C., & Chen, Y.-N. (2022). PLM-ICD: Automatic ICD coding with pretrained language models. Proceedings of the 4th Clinical Natural Language Processing Workshop. https://doi.org/10.18653/v1/2022.clinicalnlp-1.2

5. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682