[[1]](#footnote-1)

Towards Building Clinical Case Report using Machine Learning: Comparing Different PICO Based Transformer Approaches

Sabah Mohammed, Jinan Fiaidhi and Hashmath Shaik

***Abstract*—Clinical case reports generation can be automated based on innovative machine learning techniques. These reports can be triggered by the physician literature queries using the PICO standard related to unusual patterns of the complications of disease or responses to therapy. Machine learning can discover novel hypotheses which have been derived from individual cases based on similar patterns found in medical research such as PubMed. In this paper we are investigating how the new transformer models such as BERT, BioBERT and GPT can be trained and fine-tuned to identify such hypothesis which will be the cornerstone of a successful clinical case report. In this direction we are experimenting with two techniques that may help in arriving at such hypothesis based on PICO queries related to given clinical observation. The first is the summarization techniques based on the various transformers such as BERT or BioBERT that can be trained on medium sample of PubMed dataset and the large language models based on GPT-2 that are trained on a large corpus of text data to generate the hypotheses for the clinical case report. Our research paves the way on how to use the new transformer based technologies to generate clinical case reports. However, there are much of analytics and experimentations are left to our next research.**

***Keyword*— PICO questions, Evidence-Based Medicine, Clinical Case Report, Machine Learning, Transformers, Large Language Models, BERT, Bio-BERT, GPT.**

# INTRODUCTION

Compiling and publishing clinical case reports is an important evidence-based medicine (EBM) process to describe unusual patient cases and to answer clinical questions that are not generally known in the medical community practice. Actually, EBM is centered on identifying the best clinical evidence available both from the clinical practice at the bedside and the scientific evidence from the bench [1]. The process of identifying these evidences for the purpose of compiling a clinical case report involves an observation from the clinician based on sound clinical questioning protocol followed by iterative interrogation to clinical sources and medical literature as well as summarizing the evidences either manually by physician or through assistive summarization or text generation methods. The clinical question needs to address the patient case main investigation issues to seek relevant materials and build the hypotheses based on the focused search. Actually, a well-built clinical question needs to include import keywords and phrases relevant to the clinical case problem or its population effect as well as any information related to possible interventions, comparison of alternative treatments/therapies and the outcome of the used treatments/therapies methods. In this direction, the medical practice often uses the PICO standard to define the key information needed for formulating a well-built clinical question [2]. Actually PICO is the *de facto* standard to improve the specificity and conceptual clarity of clinical questions [3]. Articulating a clinical question based on PICO uses four anatomic parts—**P**roblem/**P**opulation, **I**ntervention, **C**omparison, and **O**utcome (PICO) to facilitate searching for precise answers from open research. Figure 1 illustrates the different types of clinical questions that can be expressed with PICO. Based on identifying these components from any case description, clinician can formulate a PICO question. For example, in a clinical case where *the patient is a female* *diagnosed with SLE. Her rheumatologist suggested Plaquenil medication for her joint pain, but she is interested in alternative therapies because she heard about success a neighbor had with turmeric tea.* The PICO components for this scenario should be as follows:

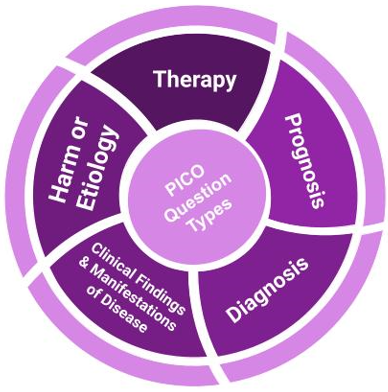
**P-** Patient with SLE

**I-** Turmeric tea

**C-** Plaquenil

**O-** Reduction of joint pain

and the PICO question can be formulated as follows: “*In adult patients with SLE, is consuming turmeric tea more effective than Plaquenil at reducing joint pain?*” The US NIH (National Institute of Health) as well as other medical institutions have developed web applications that uses PICO form to interrogate important clinical literature from many sources including the PubMed covering variety clinical repositories [4] (e.g. Clinical Trials, Randomized Controlled Trails, Systematic Reviews, Practice Guidelines, Meta-Analysis and medical Literature). Only the PubMed repository contains more than 35 million articles and is updated daily [5].

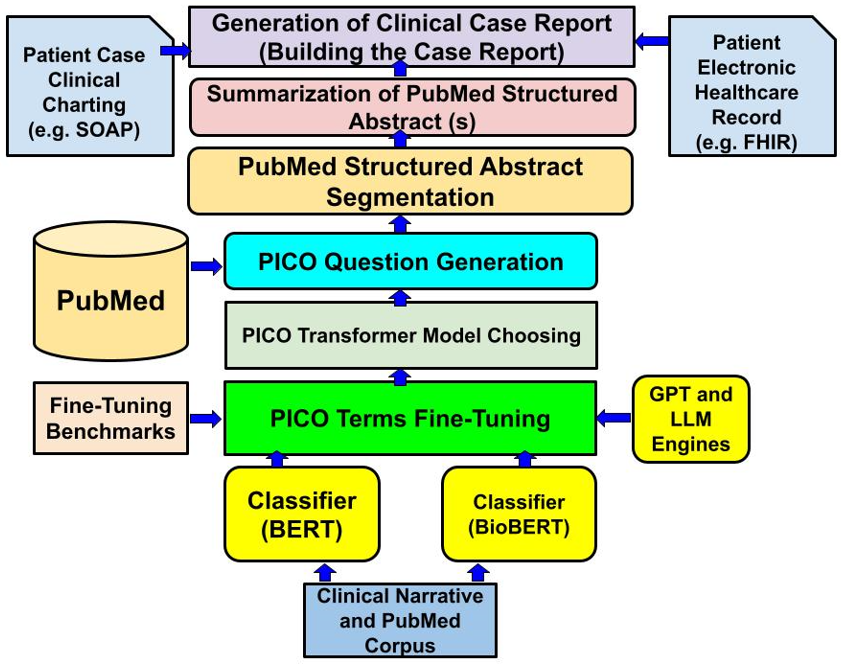


**Fig. 1:** PICO Question Types.

Although there are lots of educational tutorials on formulating PICO questions related to clinical cases, their formulation remains less straightforward [6]. For this reasons, many attempts have been introduced to automatically assist in identifying the PICO components from medical cases descriptions to help in formulating variety of PICO questions. Most important are the machine learning attempts that provided helpful results [7, 8, 9]. However, these attempts does not reveal the most effective machine learning that can identify PICO components without much of the expert intervention. In this paper we are investigating two important issues including the best machine learning technique that can be used for identifying PICO question components and how to summarize or generate a clinical case report based on forming PICO question search from sound clinical research repository like the PubMed.

# machine Learning Methods for Identifying PICO Components

As stated earlier that the fundamental skill required for practicing EBM is the asking of well-built PICO clinical questions when presented with any clinical narrative case. However, identifying automatically the PICO components from such clinical narratives require a special spider like capability involving what is called as “qualitative evidence syntheses” [10]. However, this synthesis process requires effective machine learning models that can reliably identify the relevant PICO components. However, the use of transformer models [11] (e.g. BERT, BioBERT) provides the state-of-the-art results on natural language processing (NLP) tasks such as those required for identifying PICO tags in clinical narratives [12], PubMed structured Abstract segments [13,14,15] and summarization of clinical narratives [16]. In this section we are focusing on identifying PICO terms from biomedical narratives including publications PubMed. However, the next sections build on this process of PICO term identification to recognize the structured abstract component including aim, methods, results and conclusions as well as summarizing individual and multiple structured abstracts. Figure 2 illustrates our qualitative evidence syntheses process for creating clinical case reports.



**Fig. 2:** The Framework for Generating Clinical Case Report based on PICO Questions.

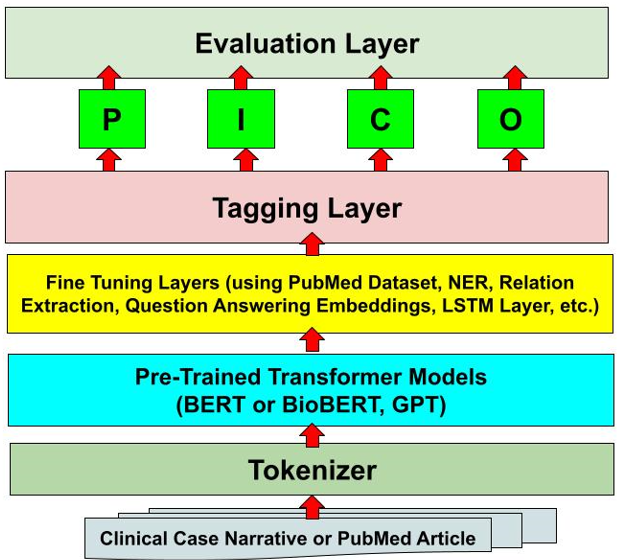
## Machine Learning Models Trained using PICO Identified Dataset using BERT and Bio-BERT

In our quest to effectively identify and extract PICO elements from clinical narratives and biomedical literature, we sought to compare the performance of two potent language models: BERT [17] (Bidirectional Encoder Representations from Transformers) and its domain-specific counterpart, BioBERT [18].

BERT, developed by Google, is a transformer-based machine learning technique for natural language processing (NLP) pre-trained using corpus of 800M words from Toronto Book Corpus and 2500M words from English Wikipedia. The BioBERT model, designed specifically for biomedical text mining using dedicated biomedical corpus that includes 4.5B words from PubMed and 13.4B words from PMC. However, BioBERT has gone through a series of improvements. The variant "*biobert-base-cased-v1.2*" was used in this paper for comparison. This variant represents an essential evolution in this sequence, introducing substantial enhancements that increase both the model's performance and its versatility. The "BioBERT-Base v1.2" extends the training corpus by incorporating an additional one million articles from PubMed. This inclusion not only increases the volume of training data, but also broadens the spectrum of biomedical sub-domains the model is exposed to, thereby refining its understanding and representation of biomedical language. However, both BERT and BioBERT are not trained to spherically identify PICO terms which make their performance in this direction lower than the expectation and necessitate the need for a fine tuning additional training [19]. For this purpose, our fine tuning started with a dataset that helps categorizing the relevant PICO words such as:

* **Aim (A)** Objective, Background, Purpose, Importance, Introduction, Aim, Rationale, Goal, Context, Hypothesis
* **Participants (P)** Population, Participant, Sample, Subject, Patient
* **Intervention (I)** Intervention, Mediation, Interposition
* **Outcome (O)** Outcome, Measure, Variable, Assessment
* **Method (M)** Method, Setting, Design, Material, Procedure, Process, Methodology
* **Results (R)** Result, Finding
* **Conclusion (C)** Conclusion, Implication, Discussion, Interpretation

This PICO tuning dataset was provided by [20] and for testing we used the tutorial by Juliet Gray of Southern Illinois University [21] that include 12 clinical cases annotated based on PICO relevancy words. Figure 3 illustrates our fine tuning strategy for enhancing the pre-trained BERT and BioBERT models.



**Fig. 3:** Fine Tuning Strategy for BERT, BioBERT and GPT.

A straight forward comparison based on accuracy measures indicates some advantage of BioBERT over BERT model (see Table 1). Similarly a closer comparison using different types of correlation with human identified PICO terms gives BioBERT more relevancies over BERT (see Table 2). However, both Tabel 1 and Table 2 show that accuracy and correlations are very weak to adopt any of the two models directly. However, research demonstrates that transformer models can be fine tuned to enhance their performance relevant to specific task like PICO questions [22,23]. For this purpose, one can use a pertained transformer model to be fine tuned you on a dataset specific to PICO search tasks. Algorithms 1 and 2 illustrates our fine-tuning to both BERT and BioBERT.

**Table 1:** Comparing BERT and BioBERT using Accuracy Measures.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **precision** | **Recall** | **F1-scores** |
| **Bert** | 0.352631 | 0.14285 | 0.0833 | 0.105263 |
| **Bio-BERT** | 0.401739 | 0.30434 | 0.0917 | 0.200579 |

**Table 2:** Comparing BERT and BioBERT using Multiple Correlation Measures.

|  |  |  |
| --- | --- | --- |
| **Model** | **Bert** | **Bio-BERT** |
| **MCC** | 0.06712417476562 | 0.3749306775111925 |
| **Cohen’s Kappa** | 0.03856690471105 | 0.2977099236641222 |
| **Pearson** | 0.02156220306095 | 0.4758343846465464 |
| **Spearman’s Rank** | 0.04054318744115 | 0.5896779102419609 |
| **Kendall Rank** | 0.03720680238634 | 0.5431688020354455 |
| **Point Biseral** | 0.02156220306095 | 0.4758343846465464 |

A black and white text on a white background

Description automatically generated

A black and white text on a white background

Description automatically generated

The dataset used in fine-tuning both transformers was introduced by Jin, Di, and Peter Szolovits [24], which is a collection of biomedical literature from PubMed searched using PICO elements. This dataset was designed to utilize a Long Short-Term Memory (LSTM) neural network-based model for the automatic identification of PICO elements in a medical text. Each PICO entity in the dataset is tagged, making it an invaluable tool for training machine learning algorithms to detect and classify such elements in biomedical literature. In terms of structure, the dataset contains original abstracts encapsulated in the 'structured\_abstracts\_PICO' file. Each record begins with a PMID, followed by the original section heading, an assigned gold label for train and test, and the section content. This information is separated by the symbol '|'. The assigned gold label is created based on a keyword check in the section heading. Moreover, an alternate version of this dataset has each section content sentence-split using the Stanford CoreNLP toolkit. This arrangement ensures that each line contains only one sentence, and all numeric values are replaced by '@'. The dataset, upon being loaded into a pandas data frame, is observed to consist of 344,636 records across five columns. The initial distribution of PICO entities is skewed, with 'O' or non-PICO elements accounting for the majority of the dataset. In contrast, the other PICO elements ('I-INTERVENTION', 'I-OUTCOME', 'I-PARTICIPANT', and 'I-POPULATION') are underrepresented. Upon an initial examination of the dataset, we identified a significant class imbalance problem, with the 'O' or non-PICO elements considerably outnumbering the other PICO elements ('I-INTERVENTION', 'I-OUTCOME', 'I-PARTICIPANT', and 'I-POPULATION'). Figure 4 illustrates the imbalance in the PICO terms at the dataset. This imbalance posed a potential issue for the performance of the machine learning models, as they could easily develop a bias towards the majority class, leading to less than optimal performance on the underrepresented classes.

A graph with blue bars

Description automatically generated

**Fig.4:** Di and Szolovits PICO Terms Dataset Imbalance.

Addressing this imbalance is crucial, as machine learning models trained on imbalanced datasets could develop a bias towards the majority class, leading to poor generalization performance on underrepresented classes. Particularly in a medical context, such as ours, where every false negative (failure to correctly identify a PICO element) could have serious ramifications, it is imperative to accurately classify all classes. To achieve a more balanced class distribution, we employed an under sampling technique. This approach reduces the instances of the majority class to match the size of the minority classes. While this method does risk discarding potentially useful information from the majority class, we considered it a necessary trade-off to mitigate bias and improve the model's performance across all PICO entities. The under sampling resulted in each PICO entity being represented 1,538 times, establishing a more balanced dataset for further analysis and model training. Given the significance of correctly identifying all PICO elements in a medical context, it was critical to address this class imbalance. The solution we employed was an under sampling technique that reduced the instances of the majority class to match the size of the minority classes. Although this method potentially discarded information from the majority class, it was a necessary trade-off to decrease bias and increase the machine learning models' performance across all PICO entities.

Upon implementing the under sampling strategy, we observed a noticeable enhancement in model performance for both BERT and BioBERT. The reduction in bias towards the majority class substantially improved the identification of PICO terms previously underrepresented. Both models' precision, recall, and F1 scores for these entities saw a significant boost, demonstrating the efficacy of this technique (see Table 3). However, we did not stop at rebalancing the classes. The next stage involved fine-tuning the pre-existing BERT and BioBERT models on the PICO-identified data. This process was aimed at enabling the models to better recognize and extract the relevant PICO terms. The fine-tuning resulted in improved performance for both models. We observed substantial increases in accuracy, precision, recall, and F1 scores. These results confirmed that both the under sampling technique and the subsequent fine-tuning significantly improved the models' ability to extract PICO components from clinical narratives.

**Table 3:** Comparing BERT and BioBERT using Accuracy Measures with fine-tuning techniques after training on balanced datasets.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **precision** | **Recall** | **F1-scores** |
| **Bert** | 0.822093 | 0.54545 | 0.6 | 0.571428 |
| **Bio-BERT** | 0.883754 | 0.88228 | 0.8837 | 0.879824 |

**Table 4:** Comparing BERT and BioBERT using multiple correlation measures with fine-tuning techniques after training on balanced datasets.

|  |  |  |
| --- | --- | --- |
| **Model** | **Bert** | **Bio-BERT** |
| **MCC** | 0.17017777230476 | 0.6495651646675051 |
| **Cohen’s Kappa** | 0.37493067751119 | 0.0541422048271362 |
| **Pearson** | -0.05916987686 | 0.2100694374718645 |
| **Spearman’s Rank** | -0.03725695418437 | 0.1560914778949903 |
| **Kendall Rank** | -0.03181491634683 | 0.1390629056129778 |
| **Point Biseral** | -0.05916987686019 | 0.2100694374718645 |

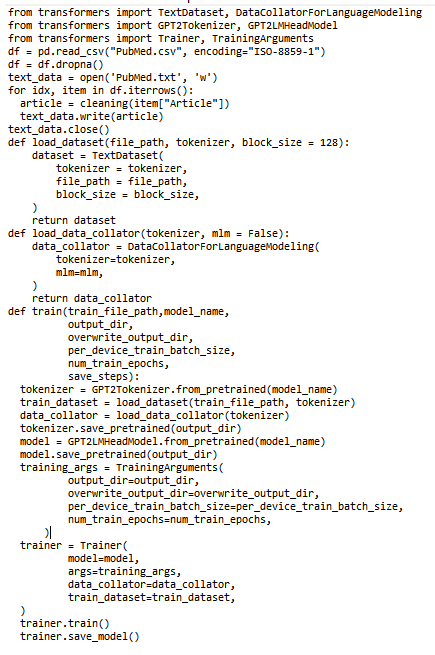
While the performance enhancements were notable, it is important to emphasize that there is still ample room for further advancements. The complex nature of clinical narratives and the inherent challenge of extracting specific PICO components suggest that there will always be a demand for better performance from these models. Future research should focus on refining these techniques and exploring other avenues to enhance the extraction and identification of PICO elements from clinical narratives.

## Machine Learning Models that Utilizes Large Language Models such as GPT

Although the transformer models such as BERT and BioBERT has over performs other convolutional neural network models like RNN by allowing parallel computing and using self-attention mechanisms to capture the relationship between key-value stores and positional encoding to improve performance, they fail to capture massively number of parameters in their model representation. In this direction, new type of large language models (LLM) started to appear in the literature since 2018 such as GPT-1 (with 117 million parameter), GPT-2 (with 1.5 billion parameters) and GPT-3 (with 175 billion parameters) [25, 26] that employs massively large transformers. The GPT-2 model was chosen for the task of identifying the PICO terms in large PubMed search due to its unique architectural characteristics compared to the other OpenAI GPT models. GPT-2 replaces the static word embeddings produced by models like BERT with more contextualized word representations which has yielded significant improvements on many NLP tasks [27].

Before training, the GPT-2 model was initialized with weights pre-trained on a large corpus of text using the BookCorpus dataset consisting of the text of around 11,000 unpublished books scraped from the Internet [28]. The use of these pre-trained weights allows the model to leverage general language understanding learned from a wide variety of texts, serving as a robust starting point for further fine-tuning on our specific task. The GPT-2 model was then fine-tuned on our balanced dataset, which was prepared through undersampling to mitigate class imbalance. The model was trained to predict the correct PICO label ('O', 'I-INTERVENTION', 'I-OUTCOME', 'I-PARTICIPANT', or 'I-POPULATION') for each token in a sentence. The model learns to understand the context and semantics of the sentence, allowing it to classify each token accurately.

To prevent over fitting, we used a split of the data into training, validation, and testing sets, following an 8:1:1 ratio. The model was trained on the training data, and the performance was periodically evaluated on the validation set. The test set was reserved for the final evaluation of the model. Figure 5 lists our Python GPT-2 training on PubMed articles based on PICO search for contextualized articles related to Asthma and Diabetics searches (PubMed.CSV). The use of the GPT-2 model for PICO element classification represents an advanced application of transformer models in the realm of biomedical literature analysis.



**Fig. 5:** Fine Tuning GPT-2 to Contexualized PubMed Searches.

Moreover, table 5 illustrates the improved accuracy of identifying PICO terms when using the GPT-2 fine tuned model is used compared to the other transformer models (e.g. BERT, BioBERT). Table 7 validate this performance improvement finding through the higher correlations found between the automatic identification of PICO terms against human identified sample.

**Table 6:** Accuracy Measures of GPT-2 with fine-tuning techniques after training on balanced datasets.

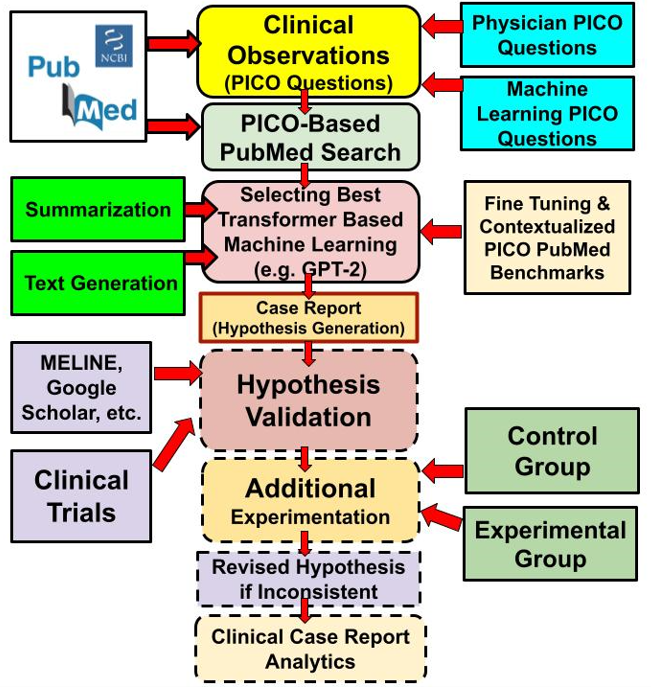
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **precision** | **Recall** | **F1-scores** |
| **GPT-2** | 0.904 | 0.90317 | 0.904 | 0.903366 |

**Table 6:** Comparing GPT-2 using multiple correlation measures with fine-tuning techniques after training on balanced datasets.

|  |  |  |
| --- | --- | --- |
| **Model** | **GPT-2** |  |
| **MCC** | 0.6137269544 |  |
| **Cohen’s Kappa** | 0.6740370511218559 |  |
| **Pearson** | 0.7375237269542 |  |
| **Spearman’s Rank** | 0.657678144 |  |
| **Kendall Rank** | 0.6375251137269544 |  |
| **Point Biseral** | 0.6269542 |  |

# Synthesizing Clinical Case Report: GPT-2 Summarization vs GPT-2 Text Generation

Clinicians make progress by strengthening their evidence based practice. Usually this process starts with an odd or unusual clinical observation that requires further investigation. Based on this observation, the clinician starts seeking further evidence from sound medical literature like PubMed to build an initial hypothesis that could point to a novel clinical case which may an attention from the medical community including conducting further investigations. In general and according to the literature, a textual clinical case report can be generated based on transformers technologies using either text summarization methods (e.g. Abstractive, Extractive) [29] or generative language methods [30]. However, nothing in the literature experimented with how such clinical report can be generated in reality based on PICO search from clinical research repositories like PubMed. This research is an attempt to generate clinical case reports based on effective transformers technologies like GPT. Figure 6 illustrates a bird view to our suggested strategy to generate a clinical case report from PICO searches including the one generated by the physician PICO queries.



**Fig. 6:** The Process of Generating Clinical Case Report and Beyond.

.

In our previous sections we found that GPT transformer technologies like GPT-2 outperform other basic transformer technologies like BERT and BioBERT. For this reason, we will focus our attention to building case reports based on GPT technology and will compare it to the other transformer technology like BERT. For the this, we will need to use the Huggingface’s transformers library (<https://huggingface.co/docs/transformers/index>) along with our PICO pertained GPT-2 and BERT models by using the *from\_pretrained()* method of the transformer class. For using the summarization methods using the GPT-2 for synthesizing a case report summary from the resulting PICO searches, we followed the four basic steps:

(1) Importing GPT-2 model and tokenizer

from transformers import GPT2Tokenizer,GPT2LMHeadModel

(2) Instantiating the model and tokenizer with gpt-2

tokenizer=GPT2Tokenizer.from\_pretrained('gpt2')

model=GPT2LMHeadModel.from\_pretrained('gpt2')

(3) Encoding the case report text to get the PICO input ids & pass them to model.generate()

inputs=tokenizer.batch\_encode\_plus([original\_text],return\_tensors='pt',max\_length=512)

summary\_ids=model.generate(inputs['input\_ids'],early\_stopping=True)

The summary\_ids contains the sequence of ids corresponding to the text summary.

(4) Decoding and printing the case report summary

GPT\_summary=tokenizer.decode(summary\_ids[0],skip\_special\_tokens=True)

print(GPT\_summary)

However, testing the accuracy of summarization requires using the ROUGE scores [31, 32] which can provide scores on the similarity between the predicted summary and original PubMed articles or their structured abstracts on the following levels:

* "**rouge1**": unigram similarity score (1-gram)
* "**rouge2**": bigram similarity score (2-gram)
* "**rougeL**": similarity based on the longest common subsequence.
* "**rougeLSum**": similarity based on the splitting the text using "\n"

The dataset employed to compile a pre-trained models for GPT-2 and BERT was derived from the TensorFlow Datasets (TFDS) collection [33], specifically from the ‘**tfds://scientific\_papers/pubmed’** endpoint. This dataset, designed for tasks related to scientific papers, offers comprehensive access to biomedical literature sourced from PubMed. The ’**tfds://scientific\_papers/pubmed’** dataset encompasses a wide spectrum of research articles across diverse biomedical domains. Each entry within the dataset is structured to provide the full text of a research article alongside its corresponding abstract. These structured abstracts offer succinct overviews of the primary research outcomes, objectives and methodologies. For our experimental framework, a subset comprising 200 research articles was extracted from this dataset. This selection was carried out to ensure a manageable volume for detailed analysis while retaining a diverse representation of biomedical subjects. Furthermore, the paired nature of texts and abstracts within the dataset facilitates a direct comparison between machine-generated summaries and author-crafted abstracts. Algorithms 3 and 4 illustrate the summarization process for the PubMed auricles using the BERT and GPT-2 technologies.

A screenshot of a computer program

Description automatically generated

A screenshot of a document

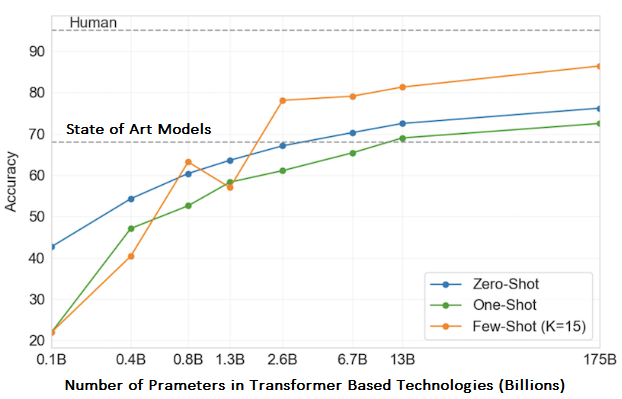
Description automatically generated

Tables 7 illustrate the summarization accuracies for PICO searches on PubMed articles based on both GPT-2 and BERT technologies.

**Table 7:** Comparing BERT and GPT-2 generated summary for a single PubMed article using Rouge metrics

|  |  |  |
| --- | --- | --- |
| **Model** | **Bert** | **GPT-2** |
| **ROUGE1** | Precision: 0.2681  Recall: 0.6727  F1 Score: 0.3834 | Precision: 0.4606  Recall: 0.4419  F1 Score: 0.4510 |
| **ROUGE2** | Precision: 0.1235  Recall: 0.3110  F1 Score: 0.1768 | Precision: 0.1585  Recall: 0.1520  F1 Score: 0.1552 |
| **ROUGEL** | Precision: 0.1208  Recall: 0.3030  F1 Score: 0.1727 | Precision: 0.1818  Recall: 0.1744  F1 Score: 0.1780 |
| **ROUGELSUM** | Precision: 0.1208  Recall: 0.3030  F1 Score: 0.1727 | Precision: 0.1818  Recall: 0.1744  F1 Score: 0.1780 |

Both technologies produce rather similar accuracies with some advantage of GPT-2 over BERT. However, it is important to note that the accuracies of the transformer models increases with number of parameters and the amount of fine-tuning data (e.g. Zero-Shot, One-Shot or Few-Shot fine-tuning) as well as the size of the training data [34, 35]. Figure 7 summarizes a comparison between various text summarization technologies that uses the transformer mode.



**Fig. 7:** Transformer Based Summarization Technologies Comparison.

However, for synthesizing the clinical report based on the generative language methods, we are conducting more research using auto-regressive language generation models that are constructed based on the assumption that the probability distribution of a word sequence can be decomposed into the product of conditional next word distributions [36]. The results of using these auto-regressive models will left to our next research.

# Conclusion

This research paper illustrates our investigations on using transformers for clinical case report synthesis. This investigation focuses on three parts. The first part uses the Bidirectional Encoder Representations from Transformers (BERT) variants (e.g. BERT Base and the BioBERT) with 12 bidirectional encoders totaling 110 million parameters and the second based on Generative Pre-Training (GPT) large language models (e.g. GPT, GPT-2, GPT-3) with 12-headed Transformer decoder (no encoders) totaling 1.5 billion parameters for GPT-2 for example. The third part is focused on using the best transformer technology to generate PICO based case report. In this direction, GPT-2 is found to outperform both the BERT and BioBERT both on the accuracy of PICO terms identifications as well as with the correlation metrics compared to human PICO terms identifications. Moreover, we investigated if text summarization techniques based on PICO-fine-tuned training that uses BERT or BioBERT can produce meaningful clinical case reports or we may need to investigate additional language generative methods based on auto-regressive models.

Acknowledgment

The first and third authors acknowledge the financial support to this research project from MTACS Accelerates Grant (IT22305-2020) and the first author NSERC DDG Grant (DDG-2021-00014).

References

1. Burns, Patricia B., and Kevin C. Chung. "Developing good clinical questions and finding the best evidence to answer those questions." Plastic and reconstructive surgery 126, no. 2 (2010): 613.
2. Huang, Xiaoli, Jimmy Lin, and Dina Demner-Fushman. "Evaluation of PICO as a knowledge representation for clinical questions." In AMIA annual symposium proceedings, vol. 2006, p. 359. American Medical Informatics Association, 2006.
3. Cheng, Grace YT. "A study of clinical questions posed by hospital clinicians." Journal of the Medical Library Association 92, no. 4 (2004): 445.
4. Brown, David. "A review of the PubMed PICO tool: using evidence-based practice in health education." Health Promotion Practice 21, no. 4 (2020): 496-498.
5. Ebbert, Jon O., Denise M. Dupras, and Patricia J. Erwin. "Searching the medical literature using PubMed: a tutorial." In Mayo Clinic Proceedings, vol. 78, no. 1, pp. 87-91. Elsevier, 2003.
6. Huang, Ke-Chun, I-Jen Chiang, Furen Xiao, Chun-Chih Liao, Charles Chih-Ho Liu, and Jau-Min Wong. "PICO element detection in medical text without metadata: Are first sentences enough?." Journal of biomedical informatics 46, no. 5 (2013): 940-946.
7. Boudin, Florian, Jian-Yun Nie, Joan C. Bartlett, Roland Grad, Pierre Pluye, and Martin Dawes. "Combining classifiers for robust PICO element detection." BMC medical informatics and decision making 10, no. 1 (2010): 1-6.
8. Chabou, S., and M. Iglewski. "PICO Extraction by combining the robustness of machine-learning methods with the rule-based methods." In 2015 World Congress on Information Technology and Computer Applications (WCITCA), pp. 1-4. Ieee, 2015.
9. Huang, Ke-Chun, Charles Chih-Ho Liu, Shung-Shiang Yang, Furen Xiao, Jau-Min Wong, Chun-Chih Liao, and I-Jen Chiang. "Classification of PICO elements by text features systematically extracted from PubMed abstracts." In 2011 IEEE International Conference on Granular Computing, pp. 279-283. IEEE, 2011.
10. Noyes, Jane. "Never mind the qualitative feel the depth! The evolving role of qualitative research in Cochrane intervention reviews." Journal of Research in Nursing 15, no. 6 (2010): 525-534.
11. Li, Yikuan, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. "A comparative study of pretrained language models for long clinical text." Journal of the American Medical Informatics Association 30, no. 2 (2023): 340-347.
12. Wang, Qianying, Jing Liao, Mirella Lapata, and Malcolm Macleod. "PICO entity extraction for preclinical animal literature." Systematic Reviews 11, no. 1 (2022): 1-12.
13. Li, Jie, Gaihong Yu, and Zhixiong Zhang. "RCMR 280k: Refined Corpus for Move Recognition Based on PubMed." Data Intelligence,MIT Press Direct, April 27 2023.pp 1-23, https://doi.org/10.1162/dint\_a\_00214
14. Hu, Yan, Yong Chen, and Hua Xu. "Towards More Generalizable and Accurate Sentence Classification in Medical Abstracts with Less Data." Springer Journal of Health Informatics Research, (2022).https://doi.org/10.21203/rs.3.rs-2026270/v1
15. Hashimoto, Kai, and Ushio Inoue. "Automatic Generation of Structured Abstracts from Research Papers by using Deep Learning." In 2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 424-429. IEEE, 2020.
16. Soleimani, Amir, Vassilina Nikoulina, Benoit Favre, and Salah Ait-Mokhtar. "Zero-Shot Aspect-Based Scientific Document Summarization using Self-Supervised Pre-training." In Proceedings of the 21st Workshop on Biomedical Language Processing, pp. 49-62. 2022.
17. Turchin, Alexander, Stanislav Masharsky, and Marinka Zitnik. "Comparison of BERT implementations for natural language processing of narrative medical documents." Informatics in Medicine Unlocked 36 (2023): 101139.
18. Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36, no. 4 (2020): 1234-1240.
19. Su, Peng, and K. Vijay-Shanker. "Investigation of improving the pre-training and fine-tuning of BERT model for biomedical relation extraction." BMC bioinformatics 23, no. 1 (2022): 120.
20. Jin, Di, and Peter Szolovits. "Pico element detection in medical text via long short-term memory neural networks." In Proceedings of the BioNLP 2018 workshop, pp. 67-75. 2018. https://github.com/jind11/PubMed-PICO-Detection
21. Juliet Gray, Picking a PICO, Online Tutorial, Available Online: <https://libguides.siue.edu/c.php?g=333872&p=2244149>
22. Wang, Qianying, Jing Liao, Mirella Lapata, and Malcolm Macleod. "PICO entity extraction for preclinical animal literature." Systematic Reviews 11, no. 1 (2022): 1-12.
23. Tinn, Robert, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. "Fine-tuning large neural language models for biomedical natural language processing." Patterns 4, no. 4 (2023).
24. Jin, Di, and Peter Szolovits. "Pico element detection in medical text via long short-term memory neural networks." In Proceedings of the BioNLP 2018 workshop, pp. 67-75. 2018.
25. Fröhling, Leon, and Arkaitz Zubiaga. "Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover." PeerJ Computer Science 7 (2021): e443.
26. Li, Conglong, Minjia Zhang, and Yuxiong He. "The stability-efficiency dilemma: Investigating sequence length warmup for training GPT models." Advances in Neural Information Processing Systems 35 (2022): 26736-26750.
27. Ethayarajh, Kawin. "How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings." arXiv preprint arXiv:1909.00512 (2019).
28. Tan, Yi Chern, and L. Elisa Celis. "Assessing social and intersectional biases in contextualized word representations." Advances in neural information processing systems 32 (2019).
29. El-Kassas, Wafaa S., Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. "Automatic text summarization: A comprehensive survey." Expert systems with applications 165 (2021): 113679.
30. Ma, Kai, Miao Tian, Yongjian Tan, Xuejing Xie, and Qinjun Qiu. "What is this article about? Generative summarization with the BERT model in the geosciences domain." Earth Science Informatics (2022): 1-16.
31. Chin-Yew, Lin. "Rouge: A package for automatic evaluation of summaries." In Proceedings of the Workshop on Text Summarization Branches Out, Barcelona, Spain, July 2004, Pages 74–81, Association for Computational Linguistics.
32. Shleifer, Sam, and Alexander M. Rush. "Pre-trained summarization distillation." arXiv preprint arXiv:2010.13002 (2020).
33. Paper, David, and David Paper. "Working with TensorFlow Data." TensorFlow 2. x in the Colaboratory Cloud: An Introduction to Deep Learning on Google’s Cloud Service (2021): 47-65.
34. Bhaskar, Adithya, Alexander R. Fabbri, and Greg Durrett. "Zero-shot opinion summarization with GPT-3." arXiv preprint arXiv:2211.15914 (2022).
35. Ahmed, Toufique, and Premkumar Devanbu. "Few-shot training LLMs for project-specific code-summarization." In Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering, pp. 1-5. 2022.
36. Zhu, Tianyu, Wen Hua, Jianfeng Qu, Saeid Hosseini, and Xiaofang Zhou. "Auto-regressive extractive summarization with replacement." World Wide Web (2022): 1-24.

1. \*Research supported by NSERC and MITACS.

   **Sabah Mohammed** is a full Professor of Computer Science with the Lakehead University, Thunder Bay Ontario P7B 5E1, Canada (phone 807-343-8777; fax: 807-7667243; e-mail: mohammed@lakeheadu.ca).

   **Hashmath Shaik** is a Graduate Student with Department of Computer Science, Lakehead University, Thunder Bay, Ontario P7B 5E1, Canada (e-mail: hshaik@lakeheadu.ca).

   **Jinan Fiaidhi** is a full Professor of Computer Science with the Lakehead University, Thunder Bay Ontario P7B 5E1, Canada (e-mail: jfiaidhi@lakeheadu.ca). [↑](#footnote-ref-1)