Writing Assignment

Improving Rare Disease Diagnosis with BERT

Marcel Santoso
Student ID : 7057881

Supervisor : Dr. Kevin Kenna
Second Reviewer : Prof. dr. J.H. Veldink

Abstract

A rare disease is an illness that affects less than one in every 2,000 individuals. There are more than 6,000 recognized rare diseases in the European Union. Collectively, rare diseases affect thirty million people in the European Union. Many doctors do not have sufficient experience and knowledge to diagnose such a diverse and rare group of diseases. As a result, rare disease patients often wait for years before receiving a definite diagnosis. Electronic health records of diagnosed patients can guide the diagnosis process of current and future rare disease patients. However, extracting relevant clinical information from millions of EHRs is challenging, especially when most diagnosis information is recorded in unstructured texts. Different clinicians may use different terms to describe the same disease and symptoms. Additionally, contexts, such as negation and cues for familial history, may affect diagnosis interpretation. BERT is one of the current state-of-the-art natural language processing (NLP) models that have been shown to understand linguistic contexts and perform NLP tasks well. This review aims to explore how BERT can improve rare disease diagnosis by processing clinical notes in EHRs. The ability of BERT to learn contextualized embeddings from data helps it to identify important words for rare disease diagnoses, such as symptoms and clinical signs, reliably. Additionally, BERT can also predict the most probable diagnosis given all information recorded in clinical notes. This information can help doctors restrict their diagnosis search space and expedite the diagnosis process of rare diseases. The use of contextualized embedding also allows BERT to be trained with imperfect labels in the fine-tuning phase. This skips the need to use labeled rare disease datasets for BERT fine-tuning process. BERT shows potential to be used in diagnosis support. However, class imbalance and limited training data for certain diseases must be sorted to improve BERT performance further.

Laymen Summary

A rare disease is a health condition that affects less than one in every 2,000 individuals. It is a term used to describe a wide range of diseases, including but not limited to cancer, developmental issues, and neurological disorders. There are more than 6,000 rare diseases recognized in the European Union (EU). Each of these rare diseases may only affect a few individuals, but collectively they affect about thirty million people in the EU alone. These people often wait for years before getting a definite diagnosis due to the rarity and heterogeneity of their disease. The lack of knowledge and experience to diagnose rare diseases among doctors is one of the causes of the delay. Doctors cannot remember the symptoms of thousands of diverse diseases by heart. Doctors need a support system that can help them recognize signs of rare diseases quickly.

Electronic health records (EHRs) of patients diagnosed with rare diseases can guide future diagnosis. They contain information regarding symptoms or signs that could be associated with rare diseases. However, extracting relevant clinical information from millions of EHRs is challenging, especially when most diagnosis information is documented in unstructured clinical notes. Doctors or nurses can use abbreviations or multiple synonyms to describe the same symptoms or diseases in clinical notes. As a result, we cannot extract relevant information from EHR accurately unless we know all possible abbreviations, synonyms, and potential errors in typing. Additionally, we must differentiate correct diagnoses from negated diagnoses and family history. In other words, context matters in extracting clinical information from EHRs.

Natural language processing (NLP) allows computers to learn and understand human language. BERT is a state-of-the-art NLP model that has been shown to extract information from texts reliably. In fact, the biggest search engine on our planet uses BERT to understand our queries and return better search results. BERT is so popular and reliable because it can generate mathematical representations of words that carry contextual and usage information. This review aims to explore how BERT can improve rare disease diagnoses. The ability of BERT to understand context allows it to identify important terms for rare disease diagnoses, such as symptoms and clinical signs, reliably. Additionally, BERT can also predict the most probable diagnosis given all information recorded in clinical notes. This information can help doctors restrict their diagnosis search space and expedite the diagnosis process of rare diseases. However, it has to be noted that BERT does not predict all rare diseases equally well. Extremely rare diseases often do not have enough documents to train BERT properly. Therefore, they are more likely to be misclassified.

Introduction

Rare diseases are a diverse set of diseases that affect millions of people every year. In the European Union (EU), rare diseases are defined as illnesses that affect less than 1 in every 2,000 individuals (1). Each rare disease may affect a few individuals, but collectively rare diseases can pose a substantial challenge to society. Currently, there are more than 6,000 recognized rare diseases reported in approximately thirty million people in the EU (1). As many as 50% of these people do not have a diagnosis for their conditions at the moment (2). Diagnosing individuals with rare diseases is complicated due to limited knowledge and expertise (3,4). Many doctors may never encounter the majority of rare diseases throughout their practice. A survey of the Orphanet database, a curated database of rare diseases, revealed that 84.5% of rare diseases are present in less than one individual per 1,000,000 (3). Due to the heterogeneity and paucity of rare diseases, some patients may receive at least one misdiagnosis and wait between 4 and 5 years before getting a diagnosis (5,6). The delay may result in decreased quality of life and shortened life expectancy as care could be postponed until patients receive definite diagnoses (2,6). Assisting doctors in recognizing patterns and identifying symptoms of rare diseases can expedite diagnosis process (6). This means that patients can receive crucial treatments earlier.

Most visits to physicians generate clinical documents that reflect one's medical history. Electronic health records (EHR) contain a wealth of information regarding the physical conditions and complaints of an individual. The records may contain clinical notes, laboratory test results, images, and treatment plans. Extracting information from these records may help us identify the phenotypes associated with rare diseases. However, most information is documented in an unstructured manner (7). Clinicians use narratives to describe observed phenotypes, diagnoses, and treatment processes (7). These types of data are not directly machine-readable or -computable (7). Information from the texts must be extracted and made into structured data to be statistically tested. However, this conversion task is not straightforward either. Word extraction programs must be able to differentiate negated diagnosis or family history from the diagnosis of patients (8). In other words, the tool must capture the relationship between disease name and other linguistic components accurately to extract information from EHRs reliably.

Computers use mathematical models to learn patterns from data. This pattern-learning ability of computers can be defined as machine learning. Natural language processing (NLP) is a branch of machine learning that focuses on the analysis and understanding of human language (9). NLP models have been integrated into virtual assistants on mobile phones and search engines to understand the queries of users. Bidirectional Encoder Representation from Transformer (BERT) is a state-of-the-art NLP model (10). BERT employs an artificial neural network, called a transformer encoder, to learn the relationship between each word in a sequence and understand language (10,11). The understanding of the relationship between words can help BERT incorporate linguistic meaning into its numerical representation of words.

General NLP methods, including BERT, can be repurposed to study massive amounts of EHRs. Mining information from EHRs may improve rare disease diagnosis by assisting doctors in recognizing symptoms and patterns associated with rare diseases. However, some modifications may be necessary for the general NLP models to mine EHR data reliably. This literature study aims to explore how NLP, specifically BERT, can support rare disease diagnoses based on historical EHRs.

Overview of BERT-based method

BERT background

People typically interact with search engines using the language that they use regularly. They can make queries by writing a sentence without following any set of rules. Computers behind the services must be able to interpret the queries accurately to return the answers that users expect. BERT was created by Google Artificial Intelligence (AI) researchers to improve computer understanding of human language (10). BERT is based on an artificial neural network model called transformer encoder (10,11). Encoder learns the meaning of a word by capturing contextual information associated with it in a text (7). The context in this setting can be understood as the usage patterns and relevance of other words in a text to the interpretation of a word of interest (7).

BERT-based workflow

BERT model refers to the stacks of transformer encoder layer that learns language representation. However, the overall BERT method involves multiple data processing stages (Figure 1). The process begins by tokenizing texts or splitting sentences into individual words or sub-words (Figure 1, the second box from the bottom). The tokenization process is followed by initial embedding assignments. Embeddings are numerical representations of a word that contain information about its usage pattern (7). In the base-BERT model, token embeddings, segment embeddings, and position embeddings are summed up to create an initial numeric representation of words in a sentence. Encoder layers incorporated context into the initial embeddings to better represent the meaning of words in a text. The output of the final encoder layer can be passed to a specialized layer that performs specific tasks, such as text classification.

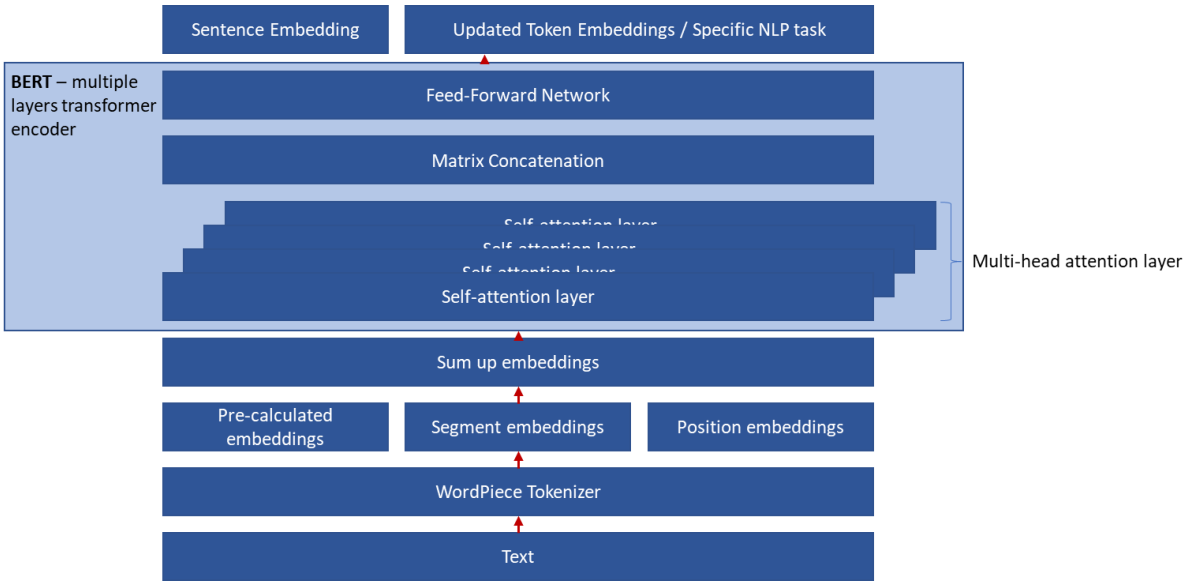

Figure 1. Overview of base-BERT method. The main steps are tokenization, initial embedding assignment, adding context to embeddings with encoders, and performing specific NLP tasks. The figure is adapted from Devlin et al. and Vaswani et al. (10,11).

BERT uses numerical embedding to represent words in a text

Computers do not process words or texts exactly as humans do. Words must be represented as numbers for computers to process. Embeddings are numerical values that represent the usage pattern and meaning of words in a text to computers (7). The process of assigning numerical representation of words begins with tokenization. Tokenization is a process of splitting sentences into smaller

functional components. BERT uses WordPiece tokenization that separates texts into sub words or morphemes, the smallest meaningful items in languages (12). Morphemes include prefixes or suffixes that can modify the meaning of a base word. BERT also adds a [CLS] token at a beginning of a text and a [SEP] token after every sentence (10). [CLS] tokens are added to contain numerical representations of the whole input text (10). The [SEP] tokens define where sentences end in the text (10). In the base-BERT method, the sum of three embeddings is used as an initial representation of a token (10). BERT uses pre-calculated WordPiece embeddings for its initial token embedding (10,12). Positional embedding represents the order of words in a sentence (10). Segment embeddings imply the border of two sentences (10). In practice, the initial embeddings can be replaced with embeddings that contain relevant information for the model purposes, such as age or gender in EHRs. The summed-up initial embeddings are updated by the encoder layers to better capture the meaning of words in the input text.

Contextualizing initial embeddings with multi-head self-attention layers

Language is complex with many components that interact together to convey information accurately. Self-attention learns the usage patterns of words in a text to infer contextual information associated with words, phrases, or sentences (7). The multi-head self-attention layers in BERT improve the context-learning process further by focusing on multiple syntactic and semantic components of a language simultaneously (14). Figure 2 shows an example of different linguistic relationships that multi-head self-attention layers can capture. The blue self-attention head was able to capture the relationship between a verb (sat) and a subject (the cat). The brown self-attention head learned identified connections between a verb (sat) and a positional phrase (on the mat). The concatenated output values from all attention heads represent the interpretation of the word "sat" in the context of "the cat sat on the mat". The ability of multi-head self-attention layers to capture different relationships at the same time allows BERT to assign contextualized embeddings that assist computers in interpreting words in a text more accurately.

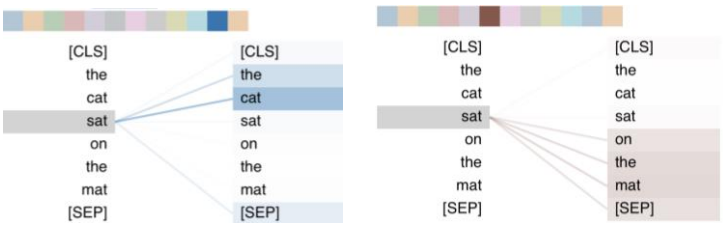

Figure 2. Different attention heads can focus on different words in a sentence. The image shows blue and brown attention head. The color intensity represents the relevance of words on the right to the word "sat" on the left. Darker color indicates greater relevance. Retrieved from (13).

Pre-training and fine-tuning of BERT models

Pre-training and fine-tuning are the two main stages in BERT implementation (10). In the pre-training stage, BERT is trained to recognize a language model and the general relationship between words. The pre-training process involves predicting masked words and sentence order (10). The fine-tuning stage allows pre-trained BERT models to perform specific NLP tasks, such as named entity recognition and text classification. This flexibility is achieved by adding a task-specific layer after the encoder stacks. The fine-tuning stage can improve the performance of BERT in various NLP tasks by changing training data and adjusting some layers in the complete model.

BERT models learn syntax and semantics associated with languages in the pre-training stage (10,14). A large amount of training data is necessary in this stage to make sure BERT captures syntax and meaning associated with a language accurately (10). Masked language model (MLM) and next

sentence prediction (NSP) tasks train BERT to capture the relationship between words and sentences (10). In the MLM training task, a fraction (commonly 15%) of tokens in the training have an 80% chance to be replaced with [MASK] token, 10% chance to be replaced with random tokens, and 10% chance to be unchanged. The MLM task trains BERT to understand word context in a bi-directional manner by predicting the masked tokens from other tokens that come after and before the masked terms(10). Bidirectionality is important as the interpretation of a word relies on contexts provided by words that come after and before it in a sentence (10). The NSP tasks are very different from MLM tasks. In NSP tasks, BERT is trained to predict if a sentence follows a selected sentence in the input text (10). This training helps BERT to understand the contexts involved between sentences.

Unlike the pre-training phase, fine-tuning phase requires labels to train a BERT model. The training data and its associated labels must be tailored to tasks that need to be performed. A task-specific layer, such as a classifier layer, can be added after the top-most encoder layer in the fine-tuning stage (10). The weights in some layers of BERT models could be adjusted to perform specific tasks. During the pre-training stage, BERT layers are specialized to identify masked tokens (15). The fine-tuning phase allows users to create a specialized BERT workflow that performs NLP tasks, other than MLM (15). For instance, fine-tuning would allow BERT to identify word categories and classify texts. Fine-tuning stage typically conserves the language model that BERT learns but adjusts the top layers that are more specialized to perform MLM tasks (15). Because pre-trained BERT models are already aware of general contexts associated with a language, BERT fine-tuning can be done with a smaller amount of data. However, fine-tuning cannot improve the performance of a pre-trained BERT model if the task involves texts from different domains (15).

Weak supervision to improve BERT training

Labels are important in fine-tuning BERT to classify texts or recognize concepts, such as diseases and disorders, in clinical narratives. Traditionally, experts label data manually to make gold-standard training or test datasets. However, obtaining such labels for millions of EHRs is impractical and time-consuming (16,17). As a compromise, weak or imperfect labels can be assigned to training examples based on a set of rules (16,17). Domain experts can determine the relevant set of rules to label texts. Additionally, medical ontology can be utilized in training BERT for concept recognition in EHRs (17). An ontology represents a word or concept with types and relationships with other concepts. Based on this relationship, words or tokens in the training dataset can be assigned a supergroup or concepts, such as diseases or organs.

Training models with imperfect labels is often referred to as weak supervision (Figure 3). The process starts with annotating texts with multiple labeling functions (Figure 3). These functions are built based on rules defined by domain experts or knowledge bases, such as ontologies. Majority vote or probabilistic model can be used to aggregate the labels. Majority vote simply selects labels that are supported by the most labeling functions. However, this approach can be erroneous because both accurate and erroneous labeling functions have equal influence (17). The use of a probabilistic aggregator would be a better option. A probabilistic aggregation function determines the most probable label by estimating the accuracy of all labeling functions in all training data points (17,18). A label is considered to be correct if it is agreed by most labeling functions (18). The probabilistic aggregation function assigns more weights to labeling functions that are correlated with the majority vote at each data point but not strongly correlated with other labeling functions (18).

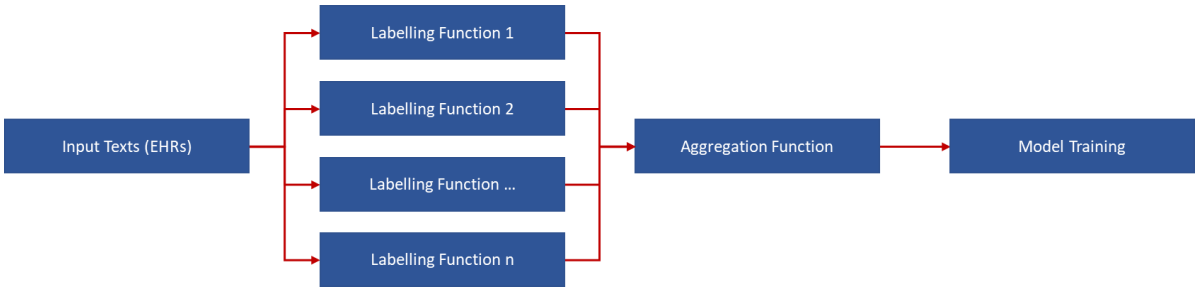

Figure 3. Weak supervision workflow. Multiple labelling functions are used to annotate text based on domain knowledge. The labels are then aggregated and used for machine learning training. Adapted from Lison et al. (16)

Labeling functions can essentially perform various NLP tasks themselves. However, this approach is undesirable as the functions may lack generalizability (18). Labeling functions that are based on heuristics, dictionaries, or domain knowledge may miss information due to erroneous spellings, abbreviations, or word choices in the text. BERT is less affected by these issues due to its use of contextualized embeddings. The embeddings that BERT assigns to words with similar meaning and usage should have similar values (10,17,18). For instance, all synonyms and abbreviations of a word should have comparable values in their embeddings. This flexibility could help BERT learn additional rules and classify more EHR documents despite the imperfect labeling functions.

Supporting rare disease diagnosis with BERT and weak supervision

The majority of clinical information of patients diagnosed with rare diseases is contained within clinical narratives. Doctors can describe symptoms, family history, and diagnosis journey more extensively with narratives than pre-defined sets of code. However, extracting relevant information from clinical narratives accurately can be challenging. Context matters in extracting information from narratives, especially when the diagnosis of rare disease patients may change. False positives can make up to 70% of rare disease queries from EHRs without taking negation and family history cues into account (8,19). The importance of context is where the ability of BERT to generate contextualized embeddings may help in improving information extraction from EHRs of rare disease patients.

Pre-trained BERT model can generate word embeddings that better capture contexts surrounding disease mentions (20). However, the numbers contained within BERT embeddings are not meaningful to doctors in clinical settings. The embeddings produced by the encoder stacks must be passed to a classifier to convert all mathematical vectors into meaningful information for physicians. The translation process can involve named entity recognition (NER) or text classification. In NER tasks, NLP models identify words or terms that convey certain concepts (7).Text classification tasks determine which rare disease is the most probable given the clinical text supplied to the model.

People use certain groups of words to describe certain concepts in communication. In supporting rare disease diagnosis, NER can be trained to identify terms that describe rare diseases, disorders, and symptoms. Doctors can then identify documents that mentions certain diseases or disorders names more easily. Segura-Bedmar et al. attempted to identify rare disease names from texts that describe diseases (20). BERT performed well with an F1 score of 0.8526 for rare disease term identification (20). The author suggested that BERT could achieve such a high F1 score due to its use of contextualized embeddings (20). This result indicates that with the right training data and labels, BERT should be able to identify rare disease mentions in text reliably.

Clinical text classification can help doctors to narrow down possible diagnoses for the patients. The classification (CLS) token is a special token in BERT that contains aggregated numerical representation of sentences or texts (10). This value can be extracted and passed to a classifier layer on top of the BERT encoder layers. Classifiers are trained to learn patterns from data and assign a label to it. In supporting rare disease detection, classifiers assign the most probable diagnosis to a clinical text based on its CLS embeddings. Li et al. used BERT with a classifier layer to predict rare diseases from two sets of clinical texts written in Chinese (21). The model had an F1 score of 0.9 when there were about 200 training documents for a disease (21). However, the F1 score dropped below 0.6 when there were less than 50 training examples available for a disease (21). The result indicates that BERT with a classifier layer can reliably diagnose rare diseases from clinical notes given that the model is trained with enough data.

Labels are necessary for fine-tuning BERT models to recognize rare disease terms or predict the most probable diagnoses from input clinical notes. When manual labels are not available for training, then weak labels can be assigned to the training data for rare disease prediction or recognition (18). Multiple labeling functions can be created with different rules to maximize the amount of labeled training data. In addition to expert opinions, knowledge bases can be used as a source of labeling rules. For instance, mentions of symptoms reported on Orphanet could be a sign of a rare disease (22). A study used BERT with weak supervision to identify terms that describe COVID19-related disorders from 2,500 clinical notes (17). The author labels disorder terms based on dictionaries, regular expression, heuristics, and terms mentioned in medical ontologies (17). Surprisingly, the weakly supervised BERT performed slightly better than the fully supervised BERT in recognizing disorders from clinical notes. The F1 score for weakly supervised BERT was 1.5 points higher than that of the fully

supervised variant (17). This result indicates that the BERT training process is quite robust to imperfect labels, possibly due to its use of contextualized embeddings that give similar values to similar words. The same approach could be implementable in training BERT models to identify rare disease terms or predict diagnoses from clinical texts.

Opportunities to implement BERT for rare disease diagnosis support

Publicly available BERT models can be used as a basis to build BERT model for rare disease EHRs quickly

Training of machine learning models, such as BERT, may require a large amount of data. This requirement is difficult to fulfill in tasks involving rare diseases due to the extreme rarity of some diseases. Fortunately, contexts learned from general EHR data during pre-training can be transferred to rare disease classification tasks (10). Transfer learning allows pre-trained models to be adjusted to perform new tasks quickly with fewer training data. Less data is required because most relationship between medical terms has been learned during pre-training. Multiple pre-trained biomedical BERT models have been made available to the public. BioBERT, Med-BERT, and BEHRT are examples of biomedical-specific BERT models that perform NER or classification tasks on biomedical data well (23-25).

BioBERT was pre-trained on PubMed abstracts and PubMed Central (PMC) articles (24). Despite the similarity in architecture, BioBERT performs better than base BERT in identifying disease names from texts. BioBERT showed a 3.73 increase in F1 score compared to base-BERT in identifying disease names from the NCBI disease dataset that contains 793 PubMed abstracts with disease annotations (26). The improvement in performance results from improved context understanding of biomedical concepts, such as genes and mutations. Fine-tuning failed to improve the performance of base-BERT on the NCBI disease dataset due to differences in word distribution between biomedical texts and general texts used to pre-trained base-BERT(24,26). Changing or adding pre-training data can improve BERT performance in domain-specific texts, such as EHR.

Med-BERT was created to predict disease from, among others, diagnoses codes, recorded treatments, laboratory results, and frequencies of certain events (23). Med-BERT was pre-trained on the Cerner Health Facts database with de-identified EHRs from 600 hospitals or clinics in the United States (23). The database contains various structured data associated with patients, such as demographics, treatment records, lab results, and diagnosis codes (23). In the pre-training process, Med-BERT tried to predict a masked diagnosis code from each patient (23). The authors reported that Med-BERT was able to predict diagnosis codes with an AUC of between 79.98 to 85.18% (23). This value means that Med-BERT was able to predict disease from data well.

BERHT is a variant of BERT that uses previous diagnoses of patients to predict conditions that may arise in the next visit (25). BEHRT was trained and evaluated using care data recorded by 674 general practitioners in the United Kingdom (25). A sentence in BEHRT training is defined as all diagnoses recorded in a visit (25). It uses diagnosis code embeddings, age embeddings, positional embeddings, and segment embeddings to represent input texts (25). The model was trained to predict possible diagnoses of patients in their future visits (25). Impressively, BEHRT was able to achieve an AUC of 95.4% for the disease prediction task (25). This value shows that BEHRT is excellent in predicting future diseases from historical EHRs of a patient.

BioBERT, Med-BERT, and BEHRT performed disease recognition and disease prediction tasks well. This means that the models can recognize the concepts that are important in EHRs. In theory, these models can be fine-tuned to identify diseases or predict diagnoses from EHRs of rare disease patients. Because

they have been performing well in their benchmarks, repetition of the pre-training process should not be necessary.

Publicly available knowledge bases are usable in weak supervision

Labeling functions for weak supervision can be built based on information contained within knowledge bases (18). Biomedical resources for rare diseases, such as ontologies, are easily accessible on the internet. Human Phenotype Ontology (HPO) and Orphanet contain the symptoms or signs and names of rare diseases (22,27). This information contained within these ontologies can be used to create weak labels for BERT fine-tuning.

Genetic test for rare diseases can help monitor BERT performance

Genetic testing is recommended for patients that are suspected to suffer from rare diseases (4,5). Genetic testing is necessary because approximately 80% of rare diseases are associated with genetic alterations (4). BERT models can help identify rare disease patients who are eligible for genetic testing. The genetic test results can then confirm or contradict the diagnoses predicted by BERT or other classifiers. If the proportion of diagnosis that are supported by genetic testing drops below a threshold, then BERT should be re-trained with old and new EHRs.

Suggestions for implementing BERT in rare disease diagnosis

Solve class imbalance issue to minimize prediction bias

Previous studies on propaganda detection with BERT indicated that BERT is quite robust to class imbalance (28). In that study, 28% of the entries in the training dataset were labeled as propaganda (28). BERT was able to achieve F1 score of 0.78 without any modifications to the training data (28). However, class imbalance can be extreme in rare disease diagnoses. The frequency of a rare disease can range from less than 1 case per one million to approximately one case per 2 thousand individuals (3). The extreme class imbalance would cause BERT model to favor diseases with larger sample sizes. If the class imbalance issue is not addressed, less frequent rare diseases are more likely to be classified incorrectly.

The proportion of rare diseases in BERT training data could be made more equal by super-sampling minority labels and under-sampling majority labels. Under-sampling randomly excludes some samples with more common labels from training data (28). Super-sampling means that the training process is repeated for some samples in the minority class (28). Additional training data can also be generated for rarer diseases by data augmentation. Data augmentation can create copies of EHRs with slight modifications. The modification can include replacing words with their synonyms, reordering words in a sentence, inserting synonyms randomly, and deleting random words in the text (29). Data augmentation may introduce noise, especially in shorter sequences (29). Therefore, the number of modifications in data augmentation must be scaled with the length of a sentence (29).

Adjustment of cost functions can be an alternative to data resampling and data augmentation. Cost functions are used to assess the performance of BERT during its training process. In model training, weights in models are updated to minimize the value produced by the cost function. Cost functions can be modified to give heavier penalties when less common classes, such as rarer diseases, are classified incorrectly (28). This would force classifiers to adjust their weights and improve the classification of rare classes, achieving a similar outcome to data resampling (28).

Data imbalance can inadvertently introduce bias in classification tasks. Classifiers are inclined to assign clinical notes of exceedingly rare diseases to larger classes. There exist multiple strategies to improve BERT training when data is imbalanced. Depending on the dataset properties and tasks, a strategy might improve classifier performance more than the others. Performance metrics, such as the F1 score, should be checked to select the most appropriate approach to dealing with class imbalance.

Include structured data in the diagnosis/prediction process

Electronic health records are not limited to clinical notes written by doctors. They may contain laboratory test results, medication lists, billing codes, and other structured data. These types of data can complement the information extracted from clinical notes in predicting rare diseases from data. Combining both structured and unstructured data from EHRs may improve the predictive power of the BERT classifier layer.

Complement BERT with rule-based approach

Doctors may write measurements, such as blood pressure and heart rate, on clinical notes. Measurements can be useful predictors of rare diseases. Unfortunately, it has been reported previously that BERT struggles in generating an accurate representation of numbers in texts (30). BERT uses WordPiece tokenizer to split sentences into individual tokens or sub words (10,12). Unfortunately, the tokenizer can also split identical numbers into multiple different sub words,

resulting in inaccurate data representation (30). In this case, rule-based approach, such as regular expression, can complement BERT in extracting information from different types of data in EHRs.

Inter-institutional agreement is necessary to reduce the need for re-training

Electronic health records are often not centralized and standardized. Each institution, country, and field can have its preferred EHR program, format, and diagnosis code (31). Such fragmented data could limit the transferability of BERT and other machine learning models (31). BERT can only recognize patterns that are present in their pre-training datasets. If different institutions use different jargon or codes, then BERT accuracy will fall because it does not recognize the terms. BERT may have to be re-trained for it to learn the contexts of the new terms. Repeating the pre-training process in different institutions can be computationally costly and lengthy (10). Therefore, agreeing on a format or a set of diagnosis codes can improve the transferability of BERT and speed up BERT setup in different institutions.

Label extremely rare disease examples with disease groups instead of disease name

Rare diseases have a wide range of frequency in the population. More than 80% of rare diseases are extremely rare with less than one case in a million individuals (3). It would be difficult for BERT to perform classification tasks on these diseases due to minimal datapoint available for training. It has been shown previously that the BERT F1 score dropped below 0.5 for diseases with less than 5 training documents (21). For these diseases, it might be safer to label such cases with disease groups instead of specific disease names until more data becomes available. This approach might be able to minimize spurious disease prediction by BERT. Medical ontologies can be used as a reference as they contain parent-child relationships between disease groups and disease names. When data is insufficient for fine-grained separation, a weak label can be taken from a higher level in the hierarchy.

Regular retraining might be required for BERT to help in diagnosing novel rare diseases

Classification models are limited by labels supplied during their training. In other words, they can only assign diseases that they have seen in their training datasets. This could be an issue with rare disease diagnoses as it has been approximated that there are between 250 to 280 new rare diseases identified annually (32). The estimated number of new rare diseases discovered annually highlights the importance of regular performance monitoring and model re-training. If accuracy falls below a threshold, then BERT should be retrained from scratch.

Report multiple predicted diseases

Individuals with the same rare disease can present different symptoms (2,20). It is also possible for rare diseases to have overlapping symptoms with other illnesses (19). It is therefore recommended for BERT to report multiple diagnoses and their probabilities. Doctors can investigate the list of probable diseases further and confirm diagnoses based on their observations and knowledge.

Generate visualization

Doctors cannot easily decode the calculations behind the result given by BERT models. There could be hundreds or thousands of weights and parameters used by BERT to perform NLP tasks. The attention visualization (Figure 2) may show which tokens or words are relevant to the contextualized representation of a rare disease mention. However, it should be interpreted with care as it is merely based on the usage patterns and relationship between words in a text.

Conclusion

Many people with rare diseases must wait four to five years before receiving a diagnosis. In theory, EHRs of patients diagnosed with rare diseases can help guide the diagnoses of current and future patients. However, clinical information from EHRs cannot be retrieved easily as it is contained in unstructured clinical narratives. Computers must be able to understand contexts, such as negation and family history cues, to minimize false positives in information retrieval. BERT can be pre-trained to capture linguistics contexts with its multi-head self-attention layer and create contextualized embeddings of words. The benefit of using contextualized embeddings is apparent in the success of BERT in retrieving and predicting rare disease mentions from texts. The use of contextualized embedding also allows BERT to be trained with imperfect labels in the fine-tuning phase. This skips the need to have manually labeled rare disease datasets to fine-tune BERT. BERT shows great potential in improving rare diagnoses, especially in helping doctors retrieve relevant information from historical EHRs and reducing their diagnosis search spaces. However, class imbalance and data quantity issues may hurt the performance of BERT-based methods in identifying extremely rare diseases from clinical texts.

References

(1) European Commission and Directorate-General for Research and Innovation. Collaboration : a key to unlock the challenges of rare diseases research. : Publications Office; 2021.

(2) D'Allessio V. The long journey to a rare disease diagnosis. 2022; Available at: https://ec.europa.eu/research-and-innovation/en/horizon-magazine/long-journey-rare-disease-diagnosis. Accessed Aug 18, 2022.

(3) Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. European Journal of Human Genetics 2020 Feb;28(2):165-173.

(4) Crowe A, McAneney H, Morrison PJ, Cupples ME, McKnight AJ. A quick reference guide for rare disease: supporting rare disease management in general practice. British journal of general practice 2020 May;70(694):260-261.

(5) Marwaha S, Knowles JW, Ashley EA. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. Genome medicine 2022 Feb 28,;14(1):23.

(6) Ronicke S, Hirsch MC, Türk E, Larionov K, Tientcheu D, Wagner AD. Can a decision support system accelerate rare disease diagnosis? Evaluating the potential impact of Ada DX in a retrospective study. Orphanet Journal of Rare Diseases 2019 Mar 21,;14(1):69.

(7) Percha B. Modern Clinical Text Mining: A Guide and Review. Annual review of biomedical data science 2021 Jul 20,;4(1):165-187.

(8) Garcelon N, Neuraz A, Benoit V, Salomon R, Burgun A. Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse. Journal of the American Medical Informatics Association : JAMIA 2017 May 1,;24(3):607-613.

(9) Cambria E, White B. Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. MCI 2014 May;9(2):48-57.

(10) Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018 Oct 10,.

(11) Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. 2017 Jun 12,.

(12) Song X, Salcianu A, Song Y, Dopson D, Zhou D. Fast WordPiece Tokenization. 2020 Dec 31,.

(13) Vig J. Visualizing Attention in Transformer-Based Language Representation Models. 2019 Apr 4,.

(14) Jawahar G, Sagot B, Seddah D. What does BERT learn about the structure of language? 2019 Jul 28,.

(15) Merchant A, Rahimtoroghi E, Pavlick E, Tenney I. What Happens To BERT Embeddings During Fine-tuning? 2020 Apr 29,.

(16) Lison P, Barnes J, Hubin A. skweak: Weak Supervision Made Easy for NLP. : Association for Computational Linguistics; 2021.

(17) Fries JA, Steinberg E, Khattar S, Fleming SL, Posada J, Callahan A, et al. Ontology-driven weak supervision for clinical entity classification in electronic health records. Nature communications 2021 Apr 1,;12(1):2017.

(18) Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: rapid training data creation with weak supervision. The VLDB Journal 2019 Jul 15,;29(2-3):709-730.

(19) Garcelon N, Burgun A, Salomon R, Neuraz A. Electronic health records for the diagnosis of rare diseases. Kidney international 2020 Apr;97(4):676-686.

(20) Segura-Bedmar I, Camino-Perdones D, Guerrero-Aspizua S. Exploring deep learning methods for recognizing rare diseases and their clinical manifestations from texts. BMC bioinformatics 2022 Jul 6,;23(1):263.

(21) Li X, Yuan W, Peng D, Mei Q, Wang Y. When BERT meets Bilbo: a learning curve analysis of pretrained language model on disease classification. BMC medical informatics and decision making 2022 Apr 5,;21(Suppl 9):377.

(22) INSERM 1997. Orphanet: an online database of rare diseases and orphan drugs. Available at: http://www.orpha.net. Accessed Aug 19, 2022.

(23) Liu N, Hu Q, Xu H, Xu X, Chen M. Med-BERT: A Pretraining Framework for Medical Records Named Entity Recognition. TII 2022 Aug;18(8):5600-5608.

(24) Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15,;36(4):1234-1240.

(25) Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: Transformer for Electronic Health Records. Scientific Reports 2020 Apr 28,;10(1):7155.

(26) Doğan RI, Leaman R, Lu Z. NCBI disease corpus: A resource for disease name recognition and concept normalization. Journal of Biomedical Informatics 2014 Feb;47:1-10.

(27) Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. American journal of human genetics 2008 Nov 17,;83(5):610-615.

(28) Madabushi HT, Kochkina E, Castelle M. Cost-Sensitive BERT for Generalisable Sentence Classification with Imbalanced Data. 2020 Mar 16,.

(29) Wei J, Zou K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. 2019 Jan 30,.

(30) Rogers A, Kovaleva O, Rumshisky A. A Primer in BERTology: What We Know About How BERT Works. Transactions of the Association for Computational Linguistics 2021 Jan 1,;8:842-866.

(31) Mugisha C, Paik I. Comparison of Neural Language Modeling Pipelines for Outcome Prediction From Unstructured Medical Text Notes. Access 2022;10:16489-16498.

(32) Dawkins HJS, Draghia-Akli R, Lasko P, Lau LPL, Jonker AH, Cutillo CM, et al. Progress in Rare Diseases Research 2010–2016: An IRDiRC Perspective. Clinical and Translational Science 2018 Jan;11(1):11-20.