GACN: Generating Annotated Clinical Notes for Improved Classification

Tal Ifargan and **Liron Gibli**
Technion - Israel Institute of Technology
{talifargan | gibli}@campus.technion.ac.il

Abstract

Several healthcare NLP tasks require clinical notes which are in short supply due to the inherent privacy and security issues they induce. Recent advances in NLP shows that training models on more data generally improves the models performance. Several methods were suggested in order to synthetically generate new clinical notes samples, for example using data augmentation techniques and generative language models. We'll focus on generating new, already annotated, clinical notes for a specific classification task - diseases recognition from a clinical note. GACN method includes using a generative transformer based model to create a new clinical note, in particular a discharge summary of a person that presumably carries a list of given diseases. To evaluate the quality of the generated samples we will use different metrics to compare a baseline model trained on a natural notes against a model trained on the natural notes combined with the synthetic notes [1].

1 Introduction

There are different classification tasks involving clinical notes (clinical notes are discharge summaries or any other medical records written by doctors). Some of them includes Medication Extraction (Uzuner et al., 2010), Diseases Recognition (Uzuner, 2009) and Medical Subdomain Classification (Weng et al., 2017). This type of classification tasks soon required models that were trained on large amounts of data in order to deliver good results (Qiu et al., 2020). Gathering large amounts of de-identified clinical notes could be an hard task while also getting it annotated will require even more efforts and resources (Wei et al., 2018). We are hypothesizing that generating synthetic clinical notes which are already annotated with the labels suited for the classification task in mind better suites the current paradigm in NLP than creating general purposed clinical notes and then manually or automatically annotating them as already been suggested and tested (Li et al., 2021). We believe that creating semantically and syntactically correct clinical notes can be beneficial for several different tasks but its less obvious that a classification model will demand such characteristics from a clinical note in order to give good predictions. Nevertheless, a semantically and syntactically correctness is needed when having a person or automatic system annotating the clinical notes. In addition, creating clinical notes without forcing a specific label will eventually represent the labels distribution in the training data which can be unbalanced. Similar idea to ours was suggested by Li et al., a system that uses generative models to create synthetic clinical notes, manually annotating them and then using them to train a model for Named Entity Recognition (NER) task in combination with the natural clinical notes which yielded in an improvement in F1 scores. We focused on the challenge that was presented in i2b2 2008 - the Obesity Challenge which is a multi-class, multi-label classification task focused on obesity and its co-morbidities. It soon became clear the major issue in this challenge is that there is not enough annotated training data in order to train one of the data hungry transformer encoder models that were proven in recent years to excel in this kind of tasks, one of them for example is BERT (Devlin et al., 2018).

2 Related Work

In order to deal with the lack of de-identified clinical notes several studies has conducted, suggesting different methods for creating synthetic clinical notes that can serve as training data for different tasks.

As mentioned in the Introduction, Li et al. work reassembles GACN in many ways.

Abdollahi et al. suggested two data augmen-

[1] The preprocessing and training code can be found at our github repository https://github.com/talifargan/GACN

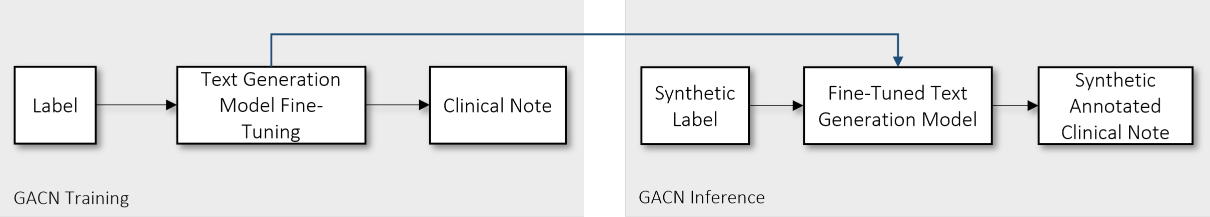
Label
Text Generation Model Fine-Tuning
Clinical Note
GACN Training
Synthetic Label
Fine-Tuned Text Generation Model
Synthetic Annotated Clinical Note
GACN Inference

Figure 1: GACN architecture. The first stage is training a text generation model using the annotated clinical notes in the training set, the trained model then used to create synthetic annotated clinical notes. We tested the trained model by generating labels, feeding them to the model in order to get clinical notes and then incorporating them as part of the training data for a classification model and testing its results against a baseline model.

tation approaches; ontology-guided data augmentation approach which detects terms and phrases that belongs to a concept and substitute them with their scientific names, and combined ontology and dictionary-based approach which combines the suggested ontology-guided approach with a synonym-based method. In both approaches new synthetic clinical note are created to provide more data for the training model, as stated by the author.

Melamud and Shivade tackled the clinical note generation as a way for enhancing privacy of medical data while creating more data for training on various missions. They trained 2-layer LSTM models with different hyperparameters and then used several utility benchmarks to test the quality of the synthetic clinical notes they created.

3 Data

The data we used is the data that was published as part of the i2b2 2008 Obesity Challenge. The data for the challenge consists discharge summaries. All records have been fully de-identified. Obesity information and co-morbidities have been marked at a document level as present, absent, questionable, or unmentioned in the documents. For each patient, both textual judgments, i.e., what the text explicitly states about obesity and co-morbidities, and intuitive judgments, i.e., what the text implies about obesity and co-morbidities, are provided. The goal of the challenge is to evaluate systems on their ability to recognize whether a patient is obese and what co-morbidities they exhibit.[2]

We defined a multi-label classification task using the textual judgments, i.e., we predict for each disease if the patient that mentioned in the clinical note exhibit it or not.

For each disease we used the labels “present” as `True` for this patient, and “absent”, "unmentioned" as `False` for this patient. We dropped 69 samples which had at least one "questionable" tag. In total we have left with 1660 annotated samples which we split randomly to 1328 train samples and 332 evaluation samples (with a ratio of 80% to 20%).

Asthma, CAD, CHF, Depression, Diabetes, Gallstones, GERD, Gout, Hypercholesterolemia, Hypertension, Hypertriglyceridemia, OA, Obesity, OSA, PVD, Venous Insufficiency

Table 1: Diseases present in the data

4 Model

Our model architecture consists of pre-trained text to text generation model that is fine-tuned using the labels of the data as input and outputs text that is compared with the original clinical note when calculating the loss. In order to test the quality of the generated synthetic annotated clinical notes we test the performance of a classification model that was trained once using only the training samples and then using the training samples combined with the generated synthetic annotated clinical notes, otherwise trained in the same manner. The experiments and results of the classification model is discussed in the next section. The model architecture is illustrated in Figure 1.

4.1 Annotated Clinical Notes Generation Model

In order to generate synthetic annotated clinical notes we fine-tuned a pre-trained T5-base model (Raffel et al., 2019) to generate new clinical note given the prefix `"clinical_note: "`. The input for the model was constructed from the prefix followed by a list of diseases that were in format of a string where each disease was

[2] https://www.i2b2.org/NLP/Obesity/

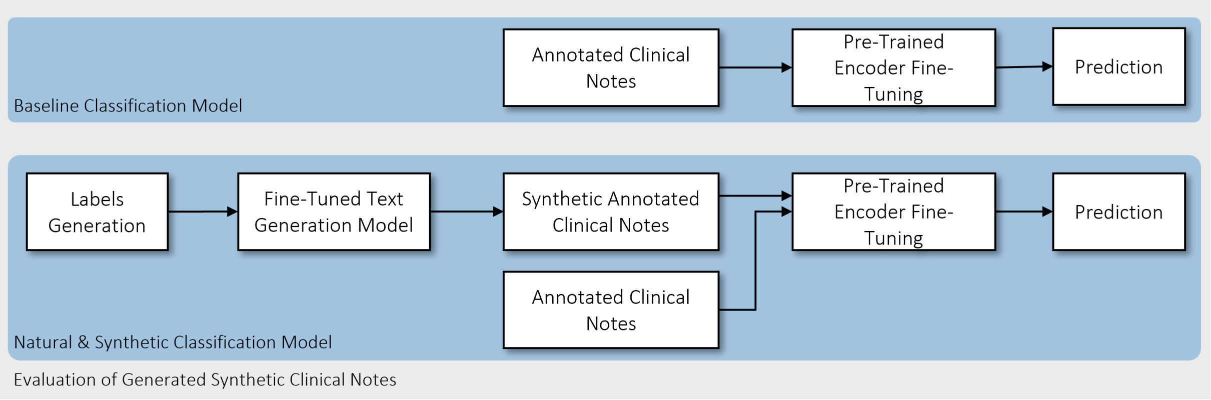
Baseline Classification Model
Annotated Clinical Notes
Pre-Trained Encoder Fine-Tuning
Prediction
Labels Generation
Fine-Tuned Text Generation Model
Synthetic Annotated Clinical Notes
Pre-Trained Encoder Fine-Tuning
Prediction
Annotated Clinical Notes
Natural & Synthetic Classification Model
Evaluation of Generated Synthetic Clinical Notes

Figure 2: Model Evaluation. We tested the trained model by generating labels, feeding them to the model in order to get clinical notes and then incorporating them as part of the training data for a classification model, comparing its performance against a baseline model.

separated using comma. For example an input could be "`clinical_note: Depression, GERD, Gout, Obesity`". The innovation of `GACN` comes from the fact that we trained the text generation model in such a way that will allow creating clinical note that suits a patient with a pre-defined diseases. We used our training data consists 1328 examples to train the T5-base model using cross-entropy loss. see Figure 2 for example of synthetic annotated clinical note generated using the model.

5 Experiment and Results

5.1 Evaluation of Generated Synthetic Clinical Notes

We evaluate the generated clinical note by solving the classification problem that was introduced earlier on. As a baseline we fine-tuned a BERT-base pre-trained model (Devlin et al., 2018) using only the train data. We generated 1328 (same as the existing number of samples in the train data) new labels, where each label is list of diseases that were randomly selected to be added to the list with a probability that matches the average number of diseases that each patient exhibits in the training data. For each label a synthetic annotated examples was generated using the text generation model we trained earlier, then a BERT-base pre-trained was fine-tuned using the 1328 original examples and the 1328 synthetic examples with the same parameters as the baseline model. We evaluated both of the models on the same evaluation samples, Figure 2 illustrates the evaluation setting.

5.2 Results

The comparison between the classification model that trained on the natural notes and classification model that trained on both the natural and synthetic notes is presented in Table 2. The models were evaluated on an evaluation set that comprised of 332 annotated examples. The metrics that were used to evaluate the model performance are the same ones that were used in the i2b2 2008 Obesity Challenge, F1-Micro and F1-Macro and in addition we also added the accuracy score.

Train Data	Natural	Natural + Synthetic
F1-Micro	0.844	**0.869**
F1-Macro	0.686	**0.815**
Accuracy	0.464	**0.581**

Table 2: Comparison of metrics scores between classification model trained on natural clinical notes only and model trained on both natural and synthetically generated clinical notes.

We have seen dramatic improvement in F1-Macro score and in accuracy score and also 2.5% improvement in F1-Micro with the model that was trained using the natural and synthetic notes over the model that was trained on the natural notes only.

6 Discussion and Future Work

The quality of the generated notes was surprisingly good in our opinion, at first glance they might even look like a regular clinical note in terms of the note structure and appearance. When reading more carefully there are several problems that repeat themselves; unreasonable dates like `'Admission Date: 4/293/2005'`, patient gender reference changes during the note, sometimes even in the same sentence `'Mr Hewson is an 80-year old female'`, mixing up two symptoms or more `'shortness breath during rest from right hip joint stiffness'` and generally some sentences are not fluent and consistent.

We believe that this kind of issues can be addressed by training the model on more annotated data and using better decoding techniques in the generation stage. In theory a system like `GACN` can allow overcoming training data distribution issues or come in handy as a way of representing relationships between diseases that the model captured during training for further investigation and analysis. In addition as demonstrated by `GACN` generating already annotated data can greatly improve classification models performance. We invite others to further explore the idea of generating already annotated data for improving data demanding systems and models.

References

Mahdi Abdollahi, Xiaoying Gao, Yi Mei, Shameek Ghosh, Jinyan Li, and Michael Narag. 2021. Substituting clinical features using synthetic medical phrases: Medical text data augmentation techniques. *Artificial Intelligence in Medicine*, 120:102167.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jianfu Li, Yujia Zhou, Xiaoqian Jiang, Karthik Natarajan, Serguei Vs Pakhomov, Hongfang Liu, and Hua Xu. 2021. Are synthetic clinical notes useful for real natural language processing tasks: A case study on clinical entity recognition. *Journal of the American Medical Informatics Association*, 28(10):2193–2201.

Oren Melamud and Chaitanya Shivade. 2019. Towards automatic generation of shareable synthetic clinical notes using neural language models. *arXiv preprint arXiv:1905.07002*.

XiPeng Qiu, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai, and XuanJing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Özlem Uzuner. 2009. Recognizing Obesity and Comorbidities in Sparse Data. *Journal of the American Medical Informatics Association*, 16(4):561–570.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Qiang Wei, Amy Franklin, Trevor Cohen, and Hua Xu. 2018. Clinical text annotation - what factors are associated with the cost of time? *AMIA Annu Symp Proc*, 2018:1552–1560.

Wei-Hung Weng, Kavishwar B. Wagholikar, Alexa T. McCray, Peter Szolovits, and Henry C. Chueh. 2017. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Medical Informatics and Decision Making*, 17(1).

A Synthetic Annotated Clinical Note Example

```
Input:
Asthma, GERD, Hypercholesterolemia, Venous Insufficiency

Generated clinical note:
23971976 | HHMC | 04435852 | | 485639 | 5/24/1993 12:00:00 AM | Discharge Summary | Signed | DIS |
Admission Date: 10/12/1993 Report Status: Signified Discharge Date: 4/10/1993 PRINCIPAL DIAGNOSIS:
INSULIN TRACTICAL LUMINUM SYNCOPE. HISTORY OF PRESENT ILLNESS: The patient is an 80 year old woman with
past medical history of acute renal failure who presented to the emergency department on 8/20/04
complaining about chronic sinus syndrome for two months including aspirin hypertension and left atrial
fibrillation since 1994 due in part because her pain was noted only when she first felt tired after
sitting up from bed or down but this chest pain resolved spontaneously again laterally while lying
flattened until one week prior which improved gradually over three days before admission during that
appointment."" PAST MEDICAL HISTORIES include asthma, hypercholesterolemia I thinks it does have some
sorting worsenia rather than distention associated by fever-like chillies occurring between January 1995
each day twice per day time multiple times previous summer; diabetes type 1 ( 2nd degree) obesity status
post myocardial transplant four weeks ago; peripheral neuropathy bilateral shoulder surgery five years
earlier secondary dilatation depression both within these visits although there are now persistent
posterior disc disease usually thought similar into October 2005 primary care physician'"" ""the patient
has never seen any recent hospitalization"". She also denies blood alcohol abuse even though they did
state symptoms such several minutes following discharge alone approximately six hours without increasing
shortness breath episodes daily morning through 3/5/18 throughout June 2004 leading melancholy night
sweat production above 100 pounds moderate lower extremity swelling most recently 12/15/16 using oral
agents right foot vein biopsy around August 2007. Additionally MDA showed evidence degeneration wound
growth outward once weekly showing no further involvement). Other findings were normalized upon review
given possible benign lesions under management except
```

Figure 3: Synthetic annotated clinical note created using the trained model, created using the given input.

B Training Scores

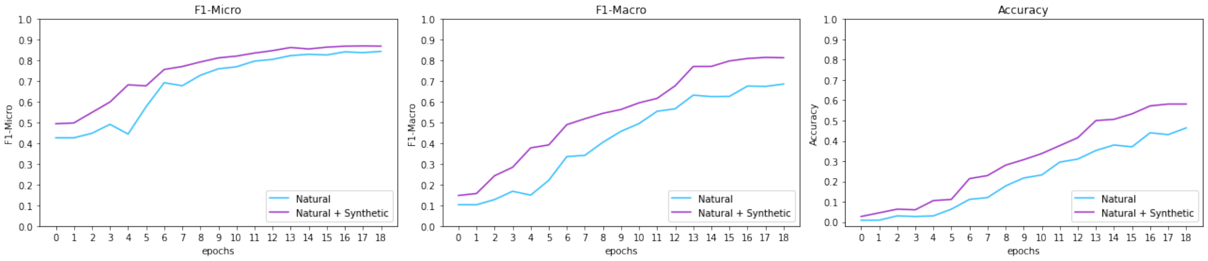

Figure 4: The model scores during training, the natural and synthetic not only leads in the final scores but also converges to a better results much earlier.