ARTICLE

Multi-task training of Transformer language models for processing radiology reports

N. Infante

INESC, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal

Abstract

Pre-trained language models based on the Transformer architecture have achieved impressive results in biomedical natural language processing tasks. One specific area that has raised interest in recent years is radiography, where textual reports are usually generated to describe findings and impressions from radiography exams. There are several practical applications related to the processing of these documents, such as automated classification, summarization or extraction of key findings. This work proposes a multi-task language model based on the Text-to-Text Transfer Transformer, commonly known as T5, that was trained on different text generation, classification, and inference tasks involving text from radiology reports. We discuss the data pre-processing and the model training strategy together with its evaluation. The proposed multi-task model achieved very good results, close to state-of-the-art results from models that were individually trained for the summarization, natural language inference, semantic text similarity, and paraphrase identification tasks. The results confirm the potential of multi-task and transfer learning for biomedical natural language processing.

Keywords: Biomedical Natural Language Processing, Radiology Reports, Transformer-Based Language Models, Multi-Task Learning, Transfer Learning

1. INTRODUCTION

The field of Natural Language Processing (NLP) has seen significantly advances in the past couple of years. The release of the paper "Attention Is All You Need" [19] has revolutionized the way we construct models and introduced a new type of neural network architecture, the Transformer, that is now vastly used. The paper demonstrated that an architecture solely based on attention layers could produce better results than the standard approaches that used recurrent neural networks.

A new generation of models based on Transformers has appeared since then. These include BERT, T5, BART, and GPT-2, which consequently opened a wide range of new possibilities regarding applications of neural models for language understanding and generation. Such models are today's standard for most tasks and applications .

One of the areas where such models have raised significant interest is the medical domain. This area raises several challenges regarding data, since acquiring large amounts of labeled medical data tends to be very hard due to the cost associated with expert labeling and the time required for such annotations. Few medical datasets are publicly available, and the quality of such data tends to vary significantly. Anonymization is also a very important topic when referring to medical data, where real patient data is being handled. Public tasks such as MEDIQA [1], contribute to reducing the shortage of training and testing data by releasing new datasets.

Radiology and radiography exams are particularly interesting in terms of possible natural language processing application. This is mostly related with the textual nature of the reports describing radiology studies. Radiology reports also present a consistent structure that is usually used to describe the impressions and findings of an exam. The practical applications of processing information on these documents can be automatic classification and labeling, summarization and extraction of key findings, or disambiguation of text. Chest X-ray is the most common imaging study performed worldwide. A large and publicly available dataset is MIMIC-CXR [10], i.e. a radiology report dataset that contains a total of 227,835 radiography studies of chest x-rays.

In this paper, we focus on exploring the application of recent language models, using the Text-to-Text Transfer model to demonstrate the capabilities of such architectures when applied to the medical domain, and radiology in particular. The decision to use this model is supported by recent literature that indicates the potential of T5, and the fact that pre-trained T5 models currently hold state-of-the-art [14] results in several medical tasks such as natural language inference.

We performed several experiments using the MIMIC-CXR dataset of radiology reports, where a multi-task training strategy was used with the following tasks: summarization, natural language inference, paraphrase identification, semantic text similarity, and classification. After fine-tuning, we evaluate and compare the performance of our model with the current state-of-the-art models [14].

Our model is based on the original T5-base and it was fine-tuned with a multi-task strategy on the tasks of natural language inference, semantic text similarity, paraphrase detection and classification. The same model was later fine-tuned on summarization, achieving comparable results with models reported in the literature for the summarization task without making any change to the original architecture. This same model was the base to another model that was fine-tuned on the all other mentioned tasks, except classification. The performance achieved for each fine-tuned model was comparable to reported results found in the literature. Overall, our multi-task models show promising results with our particular training strategy, with a drop in performance for each specific task compared to the overall state-of-the-art results, but still being

able to generalize enough to achieve good results in all tasks. As our main contribution, we provide a base for future work regarding multi-task NLP models in the medical domain, in particular the radiology area.

The rest of document is structured as follows. Section 2 we presents previous related work regarding multi-task training strategies and a description of natural language processing tasks covered on our work. Section 3 explains our proposed approach, beginning with the choice of our base model, training strategy, and data augmentation techniques. Section 4 summarizes our overall findings, including an overview of the datasets and metrics used to report our results. Finally, Section 5 presents our major conclusions and defines possible steps for future work on the area of medical NLP.

2. Related Work

Pretrained language models have become a hot research topic in the past half-decade. The promise of transfer learning by pre-training a model and later fine-tuning it is nowadays the most common approach to NLP problems [15]. Experimentation in several medical NLP tasks have been reported in the literature and here we provide a brief overview on previous methods, for different tasks.

2.1 Automatic Labeling

The current availability of medical datasets is very scarce. Annotation and labeling require experts in the field and it is very time-consuming, which creates a bottleneck in terms of the data available to train and evaluate machine learning. Attempts to create automatic labelers are not new but these in turn, also need large amounts of training data.

In the paper by Meng et al. [13], the authors focus on the problem of delays in the communication of urgent clinical findings in radiology exams. The authors trained and fine-tuned a model in a large radiology report dataset, to identify critical reports with time sensitive findings and to feed this information to an existing pipeline that delivers the information to physicians. The proposed self-supervised contextual language representation model is based on a pre-trained BERT [6]. The decision to use contextual embeddings resulted in better results than the state-of-the-art at the time, based in word2vec representations.

Another automatic labeler named ChexBERT was developed by Smit et al. [17]. ChexBERT is also a BERT-based model that was previously pre-trained in biomedical texts. The paper proposes a training scheme where the model is trained on radiology reports with annotations from a rule-based labeler named CheXpert [8], and later fine-tuned in expert annotations augmented with a back-translation strategy. The idea behind this training is to use the already learned information of the rule-based model and feed it to the new model in the form of the generated outputs. The results outperformed previous rube-based labelers on the MIMIC-CXR dataset.

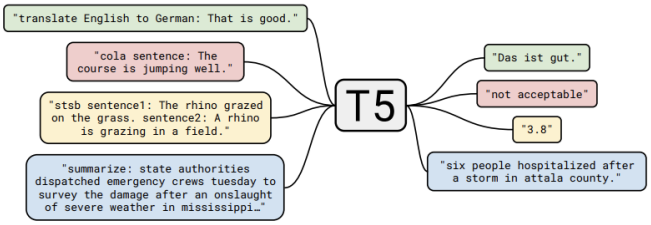

Figure 1. The text generation approach in the T5 model, together with the use of prompts.

2.2 Natural Language Inference

Natural Language Inference (NLI) is a task designed to assess the inference relation between an hypothesis and a premise expressed in natural language. Three types of inferences can be verified, entailment, neutral or contradiction.

The release of MedNLI in 2019 [16] allowed anyone to test their models in the NLI task, specifically focused on the medical domain. Several models have been specifically trained and tested in the MEDNLI dataset. For example, Phan et al. [14] proposed SciFive, i.e. a T5 based model that is pre-trained in a biomedical corpus of PubMed abtracts, followed by a training phase using a multi-task learning strategy and later fine-tuned in five biomedical NLP tasks. The paper defined new state-of-the-art scores for several tasks including also the MedNLI dataset.

2.3 Semantic Text Similarity

Semantic text similarity is an NLP task designed to quantitatively assess the semantic similarity between two text snippets [22]. The release of public tasks in the last couple of years, focused on this topic has helped the exploration of new models that perform better in medical STS.

In Wang et al. [20], BERT-based models are tested in the STS task. In particular, the authors try different ways of extending training data, using unlabeled domain data which is assigned labels from a general model. This strategy produced new state-of-the-art results in one of the datasets used.

2.4 Summarization

The task of summarization of radiology reports, although technically very challenging, presents several incentives to promote and develop advancements. In particular, the most important factor is the direct impact on the efficiency of clinical communication pipelines and the acceleration of the radiology workflow [13].

The MEDIQA shared task focused specifically in this domain. In the 2021 edition, the overview article of the task from Abacha et al. [1] reports the results presented by the participants, as well as a brief overview of the type of models that were used. From the total of 7 teams, 6 presented pre-trained models based in BART [11] or PEGASUS [23].

2.5 Medical Paraphrase Detection

Paraphrases are defined by Bhagat and Hovy [3] as sentences that convey the same meaning using different wording. Unfortunately, very few datasets are available for this task when considering medical domain data. A particularly well-known dataset is the MSRP corpus released by Microsoft [7], nonetheless featuring general text.

3. Proposed Approach

Pre-trained models allow us to reuse features from the general domain, learned from large corpora of unlabeled data, to generalize better when training and fine-tuning specific models. This approach tends to result in better performance when compared to training a model from scratch only in task-specific data. Here we propose to take advantage of the power of transfer learning, with the use of the text-to-text-transfer transformer (T5) proposed by Raffel et al. [15]. Text-to-text means that, for each input received, the model returns a string as an output, making it appropriate for question answering, summarization, and other text generation or classification tasks.

3.1 The T5 Model Architecture

The text-text transfer transformer (T5) model [15] is based on the original Transformer architecture proposed by Vaswani et al. [19], with some minor differences. The model is composed of an encoder-decoder structure. The encoder component is composed of a stack of blocks, which contain a self-attention layer followed by a feed-forward network. The decoder structure is very similar to the encoder, except that after the self-attention layer the model also includes an encoder-decoder attention layer.

The original article also presented five size versions of the T5 model. For our work, we developed all the experiments using the original T5-Base model, which is the original baseline model with roughly 220 million parameters. No changes were performed to the original architecture, training objective, and vocabulary of the original model publicly available at the HuggingFace website[a] To pre-train the original models, Raffel et al. [15] used the Colossal Clean Crawled Corpus (C4), i.e. a 750GB size corpus based on a clean version of Common Crawl's original web archive.The T5 models were pre-trained using an unsupervised denoising task that is based on masked language modeling and word dropout regularization. The original input tokens are randomly sampled and 15% are dropped. These tokens are later replaced by unique sentinel tokens. The model's objective is to predict the sentinel tokens that correspond to the original dropped out text.

3.2 Multi-Task Learning

The original paper by Raffel et al. [15], as well as other recent papers, suggest that multi-task learning can provide good results across several tasks. We use maximum likelihood as objective, together with teacher forcing to fine-tune the T5-base already pre-trained in multiple tasks, like questions-answering, summarization, or natural language inference. A great advantage of using T5 is that the model was already trained in several tasks using a prompting approach(i.e. ,at the beginning of the input, a prompt dedicated to a specific task is added, serving as a clue for the model to know which task it is handling). In most of our tasks we were able to re-purpose some of the prompts used during pre-training of the T5 baseline, by pre-appending the prompt at the beginning of the examples. Table 1 summarizes the tasks that we considered, which are also explained next.

[a] https://huggingface.co/t5-base

3.3 Fine-Tuning

We preformed fine-tuning on 5 biomedical tasks:

- Natural Language Inference (NLI);
- Semantic Text Similarity (STS);
- Paraphrase detection (PD);
- Summarization;
- Classification (CHEX);

Our training scheme is divided in two phases. On the first phase we perform a multi-task training where we fine-tuned a T5-base model with data from the all tasks mentioned above with the exception of summarization (referred to as small tasks). The second phase uses the last checkpoint from the first phase and is followed by fine-tuning the model on the summarization task. After phase two, we proceed to fine-tuning the models for each individual small task, with the exception of the classification task. Figure 2 presents an illustrative representation of our training scheme.

One of the major reasons for choosing this specific order of training is that the summarization task differs significantly in input and label size, compared to the other remaining tasks that were already mentioned. This creates a bottleneck in terms of training. To perform a true multi-task training with all tasks, we need to considered an input and label sizes that fit well the summarization task, but this requires extensive padding to the vectors for the remaining tasks. This comes from the fact that the T5 model requires all input and labels with a training batch to have the same length.

3.4 Data Augmentation

As mentioned before, biomedical datasets are very scarce and limited. The original datasets for the small tasks had two disadvantages, namely the reduced size and not being related to radiology context. To surpass these limitations, we considered several common strategies to augment the available data with radiology data (MIMIC-CXR examples) and subsequently help the model learn with more examples.

3.4.1 Back Translation

A common strategy to increase the dataset size involves the use of back-translation, where an input sequence is translated to another language and translated back again. This approach tends to modify the original sentence with paraphrasing [4]. All examples were translated to Portuguese, French, Spanish,

Table 1. Prompts appended to the input specified by task.

Task	Prompt	Model Input
Summarization	summarize	summarize sentence1: "sentence 1" sentence 2: "sentence 2" , "label"
NLI	mnli	mnli hypothesis: "hypotehsis" premise: "hypothesis" , "label"
STS	stsb	stsb sentence1: "sentence 1" sentence 2: "sentence 2" , "label"
PD	mrpc	mrpc sentence1: "sentence 1" sentence 2: "sentence 2" , "label"
Classification	chex	chexbert hypothesis: "impression" premise: "observation", "label"

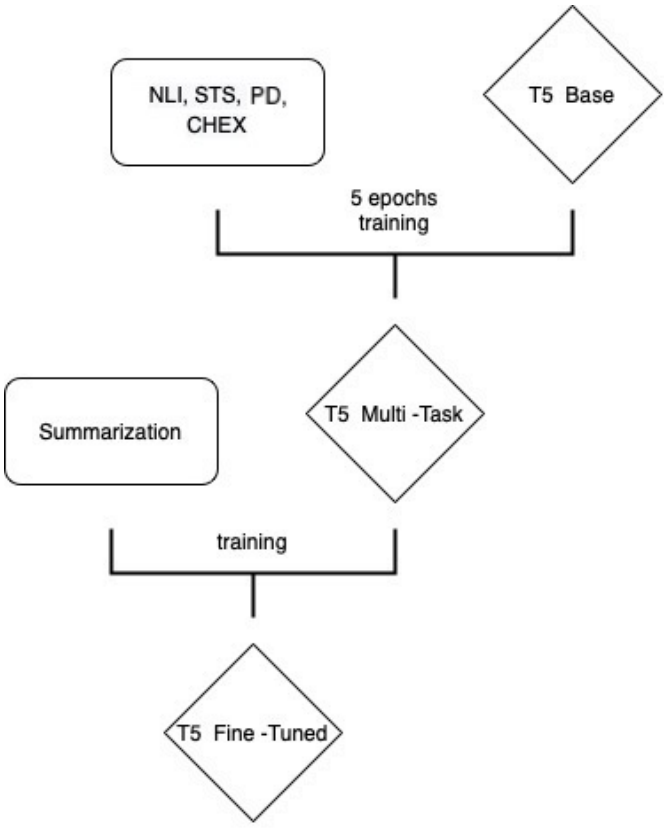

Figure 2. Training Scheme used to trained our model.NLI - Natural Language Inference, STS - Semantic Text Similarity,PD - Paraphrase Detection

and German, and later back to English. The example with the higher number of changes compared to the original English input was selected as the new example and added to the dataset. This allowed doubling the number of examples of the datasets, without introducing significant noise.

3.4.2 *Artificial Task*

We created an artificial task(i.e., the classification task) to simulate the labels given by the ChexBert labeler [17] and add an additional task to our multi-task training. We then trained a T5-base model with the impressions from MIMIC-CXR and as label the outputs given by the original ChexBert labeler. The ChexBert model is based on a pre-trained biomedical model that was fine-tuned on radiology labels from another model and fine-tuned again on a small set of expert annotations augmented with back-translations. The task of report labelling is to extract the presence of one or more cliniclally relevant observations from free-text radiology reports. This particular model performs labelling in 14 different medical observations: Pneumonia, Fracture, Consolidation, Enlarged Cardiomediastinum, No Finding, Pleural Other, Cardiomegaly, Pneumothorax, Atelectasis, Support Devices, Edema, Pleural Effusion, Lung Lesion and Lung Opacity. The labels can be blank, positive, negative or uncertain, with the exception of No Findings which can only be either blank or positive.

3.4.3 *Data Generation*

The relation between findings and impressions tends to be very strong in terms of similarity of tokens and general content. We decided to explore this fact to increase the number of examples in our datasets. Using the ChexBert Labeler to generate labels for our MIMIC-CXR findings and impression in 14 different categories, we explored this information to identify relations between both sections of the radiology reports. In particular, the impressions as a premise should be sufficient to infer the overall summary, in which case say there is an entailment relation. If one cannot infer neither contradict the hypothesis (findings section), this represents a neutral relation. Lastly, if one can use the premise (impressions sections) to deny the findings, a contraction relation is found. A threshold of similarity is defined between the 14 labels to classified each pair of findings/impression has having a relation of entailment, neutral or contradiction.

We attributed the entailment label to a pair of findings, impressions that achieved a ratio superior to 0.7 of equal labels divided by the total labels different than blank and, with zero labels that where opposites i.e. positive and negative label for the same observation on the findings impressions pair. A contradiction label was assigned when the pair of findings impressions achieved a ratio superior to 0.2 of opposite labels divided by the total labels different than blank. Th neutral label was given when the pair of findings/impressions achieved a value less than 0.5 on the same ratio described for the entailment label and conditions.

A summary of findings in radiology reports is also meant to preserve key features of the impression section, basically being a shorter version but with the same overall meaning. In other words, there is a paraphrase relation between findings and impression, since the impression section should be descriptive enough to pass the same content with less words. Additional PD data was thus created exploring this relation.

This PD dataset has two possible labels: equivalent, which

means that the sequences are equivalent to each other in meaning, and not equivalent. The following criteria was used to label the examples: If all labels of the ChexBert labeler were equal for both impressions and findings, we attribute the label equivalent. All other cases different from the previous were considered not equivalent.

4. Experimental Results

This section describes our experimental procedure and the obtained results. A full comparison with the current state-of-the-art results on the each task is provided. All experiments were run on Google Colab.

4.1 Datasets

We now provide a brief description of the medical datasets and the pre-processing techniques that were used.

4.1.1 MIMIC-CXR

The MIMIC-CXR dataset, released by Johnson et al. [10], is one of the largest radiology datasets publicly available. It corresponds to a total of 227,835 radiographic studies, particularly chest studies. Each report is composed of findings and impression sections. To sample one such big corpus, we used the script provided on the MEDIQA 2021 Radiology Report Summarization [1] task to generate a training set of 91,544 radiology reports. As evaluation set we used the MEDIQA 2021 task test set, representing a collection of 600 chest X-ray reports. No further pre-processing was performed.

4.1.2 ClinicalSTS 2019 dataset

Released for the 2019 n2c2/OHNLP task [22], this dataset was derived from electronic health records of the Mayo Clinic, combining 1068 sentence pairs from the previous year task that used a subset of MedSTS with new 574 pairs. Each pair is given a score between 0 and 5, where the first is the lowest and the last the highest score possible. A total of 1642 pairs are used for the train subset and 412 for the testing subset.

4.1.3 MEDNLI

Release in 2019 [16], the MEDNLI dataset is composed by pairs of premises and hypothesis. Each pair is associated a label of entailment, neutral or contradiction. The original premises were sampled from the MIMIC-III dataset [9] and the hypothesis were manually annotated by physicians. A total of 11,232 training examples, 1395 validation examples, and 1422 test examples are available.

4.1.4 MED PD

We used a small subset of medical paraphrases contains 150 examples for training and 60 validation examples, compiled by Aditya [2] and that was based on examples from the i2b2 (2018) cohort detection task dataset. [18].

4.1.5 Augmented Datasets

As explained under the data augmentation section, we increased the size of the original datasets with some simple techniques, like back-translation or generation of new examples through some heuristics. Table 2 presents a comparison between the original sizes of the datasets, and which augmentation strategies contribute to increase each dataset.

4.1.6 ChexData

The ChexData dataset was created for this work, specifically to train a T5 model on which to simulate the labels given by the original ChexBert labeler. The examples are composed by MIMIC-CXR impressions and their respective observation labels given by the ChexBert labeler. This dataset was only created and used for the chext X-ray labeling task.

Table 2. Original Dataset Sizes and changes due to data augmentation.

Dataset	Original Size	Back-translation	ChexBert	Final Size
PD	150	-	1476	1626
NLI	11232	11232	12000*	34464
STS	2020	2020	-	4040
ChexData	205960	-	-	205960
MIMIC-CXR	91544	91544	-	183088

* This data was augmented using back-translation.

4.2 Metrics

Taking into consideration the current metrics being used for each task and that are widely reported in the literature, this next section provides a brief description on the metrics that were used to evaluate the different tasks.

4.2.1 ROUGUE

Recall oriented understudy for gisting evaluation (ROUGE) is the standard metric used to evaluate automatic summarization [12]. The metric evaluates the quality of the produced summaries by comparing them with an human reference that is meant to serve as an example of a perfect summary. The original paper proposed 4 different ROUGE scores, but we opted to only report ROUGE-2, since it is the official metric for the MEDIQA task of radiology reports summarization [1]. ROUGE-2 measures the co-occurrence of bi-grams between the machine generated summary and the references.

$$\frac{\sum_S \sum_{ngram_n \in S} Count_{match}(ngram_n)}{\sum_S \sum_{ngram_n \in S} Count(ngram_n))} \quad (1)$$

In this equation S represents the set of summary references and $ngram_n$ is an ngram belonging to one of the references.

4.2.2 Accuracy

Several tasks that we explored correspond to classification problems, like medical inference or paraphrase detection. The common metric used to evaluate classification performance is

accuracy. We can simply define accuracy as being the number of right predictions, divided by the number of total predictions.

4.2.3 Pearson Correlation Coefficient

The task of semantic text similarity differs from all the remaining tasks by being actually a regression task. The output of this tasks, varies between a range of 0 and 5 with increments of 0.2. This is not an obstacle for the T5 model, which deals with this problem by converting the original task into a classification one, with a total of 30 possible labels. In order to make an accurate judgment of the performance of our model, we reported our results using the Pearson Correlation Coefficient, which is reported in the literature and public tasks, as a good indicator of similarity between two lists of scores [22].The following equation describes how to compute the coefficient value:

$$Pearson = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \quad (2)$$

4.2.4 F1-Score

Besides considering accuracy, the evaluation of the performance of our models on the paraphrase detection and medical inference tasks was also performed using an F1 metric, which performs a geometric mean of the per-label recall (R) and precision (P). The outputted score was unweighted, which means that that we did not take into account the distribution of the labels in our test set, due to the fact that it is assumed that all present an evenly occurrence distribution.

$$F1 = 2\frac{P \times R}{P + R} \quad (3)$$

We compute the recall by dividing true positive labels by the sum of true positive and false negative labels. The precision is calculated by dividing true positive labels by the true positive and false positive labels.

5. Experimental Results

This section presents our experimental results. The following table is a description of the most relevant models trained during our work and that are worth reporting.

Table 3. Model's Description and Training Parameters

Model	*BaseModel*	*TrainingEpochs*	*TrainDatasets*
Base	-	-	C4
X1	T5-Base	1.5	ChexData
F1	T5-Base	2.5	MEDNLI,STS and MSR
F2	T5-Base	5	MEDNLI,STS and MSR
S1	F2	2.5	MIMIC-CXR
S2	F2	5	MIMIC-CXR
S3	F2	1	MIMIC-CXR

5.1 Classification Task

We trained a T5-base model on outputs from the ChexBert labeler of the MIMIC-CXR dataset. The idea was to have a model that could accurately generate predictions of labels when compared to the original dataset. The model X1 was trained for 1.5 epochs in total on the ChexBERT dataset with a weight decay of 0.01 and a value of 500 warm-up steps. We than compared the performance of X1 using accuracy when compared to references produced by the original ChexBERT, on a test set of 20000 examples, different from the examples used during the training phase. The overall results are documented in Table 4 . To reduce the size of the dataset we decided to not predict labels for the class "No Findings".

Table 4. Acccuracy of the T5-model, for each observation category,trained on the ChexData dataset and tested a test set of 20000 random ChexBERT labels from the MIMIC-CXR dataset

Observation	Accuracy
Edema	0.9925
Fracture	0.9945
Consolidation	0.9855
Enlarged Cardiom.	0.9855
Pneumonia	0.9945
No Finding	-
Pleural Other	0.9980
Cardiomegaly	0.9945
Pneumothorax	0.9995
Atelectasis	0.9920
Support Devices	0.9875
Pleural Effusion	0.9955
Lung Lesion	0.9945
Lung Opacity	0.9765

5.2 Training Parameters

The number of epochs and conditions were adjusted, taken into account the difference in the examples between the multi-task datasets and the original single task datasets. A weight decay value of 0.01 and the value of 500 for the warming steps was used in all experiments. During the generation phase only beam search was used and the max length was adapted according to the tasks.

5.3 Multi-task Training

In order to validate our multi-task learning approach, we intended to verify that by training the model in several generation tasks we would see an improvement on the model's

performance on those same tasks. We compared training the model first individually in some tasks and than we compared the performance results on Table 5.

Table 5. Results - Small Tasks, with FF1 and FF2 correspond to the final fine-tuned model, trained on each task individually.

	PD		*STS*	*NLI*	
	F1Score	*Accuracy*	*PearsonCor.*	*F1Score*	*Accuracy*
Base	0.39583	0.63333	0.69233	0.36050	0.41421
F1	0.79107	0.81667	0.84018	0.73117	0.73769
F2	0.77815	0.81667	0.83952	0.78746	0.78832
FF1	0.81667	0.77011	0.81587	0.78551	0.78443
FF2	0.86667	0.84596	0.81490	0.79958	0.79867
IBM [21]	-	-	0.90100	-	-
SciFive [14]	-	-	-	-	0.86570

5.4 Natural Language Inference Task

The Natural Language Inference (NLI) task was used on the first phase of our training scheme to fine-tune the T5-base model, combined with the STS, PD and ChexBERT tasks. The performance of our model after 2.5 and 5 training epochs on this multi-task training scheme are reported on Table 5. We see that the model's performance on this task improved with more significantly with more training epochs, noticeable by comparing model F1 and F2. Another interesting result was the T5-base zero-shot attempt, which performed relatively bad on the NLI task. This suggests that although being pre-trained on natural language inference [15], the performance of the model is very sensitive to different domains for this task.

We also fine-tune our model on the NLI dataset after the second phase of our training scheme, this time only with this specific task. If we observe the results on Table 5, we see that after 2 epochs of fine-tuning (model FF2), the f1-score and accuracy on this tasks increases to values close to 0.80. This is an improvement compared to the performance of first phase model. Overall, the results obtained are not very far from current state-of-the-art models [14].

5.5 Semantic Text Similarity Task

Similarly to the previous tasks, the Semantic Text Similarity task was used on the first phase training scheme of our model, combine with the PD, NLI and ChexBert tasks. A striking result that can be seen on the section of the STS task on Table 5 is the relatively high correlation score achieve by the T5-base model zero shot attempt. The T5 model was pre-trained on semantic text similarity tasks [15], nevertheless, it shows a relatively strong performance for a particular radiology and medical domain that was not previously anticipated. The performance of our models after 2.5 and 5 epochs, also present some interesting. One can notice that there a decrease of the correlation score when the model is trained for more epochs (model F2 compared to F1). The model's performance is still an improvement of 0.15 (F1 model) and 0.14 (F2 model) compared to the zero-shot attempt. A possible explanation for this decrease in overall performance is most likely due to the model having reached a point of overffiting, a typical phenomenon when the model gets stuck to the training examples and loses its capability of generalization, leading to poorer results when reaching a stage of inference to predict labels.

A second fine-tuning with only this task, was performed after the second phase of our training scheme. The results presented on Table 5 show that we it was not possible to achieve better performance than the first training phase. This is most likely due to the nature of the classification task that perhaps requires more training data, so that it does not overfit so easily. Secondly, we conclude that this task doesn't appear to benefit from the second phase training.

5.6 Paraphrase Detection Task

Regarding the Paraphase Detection task, similarly to the previous tasks, it was also used on the first training phase of our scheme, combined with the STS, NLI and ChexBERT tasks. All results of the first training phase are presented on Table 5. Similarly to the NLI tasks, we see that the zero-shot attempt of the T5-base model produces very poor results despite being pre-trained on paraphrase corpus [15]. The same phenomenon that occurred for the STS task happens for the PD performance, where we see a decrease on the F1-score when the number of training epochs is increased. For this specific dataset there is no available literature for comparison, but we still opted to include it in order to increase the available training data of our multi-task approach.

The second fine-tuning with this task only is displayed on Table 5 it reports a great improvement compared to the first training phase results.

5.7 Summarization Task

When it comes to the summarization task, the results are reported on Tables 6, 7 and 8, under the models S1, S2 and S3, and for different metrics. This task is only used on the second training phase, where we fine-tuned the T5-base model fine-tuned from the small tasks. The T5-model was already pre-trained on the summarization task [15], but still achieves poor results similar to the previous tasks reported when we perform a zero-shot attempt.

The results achieved after fine-tuning the previous multi-task model show great improvements compared to the T5-base zero shot approach. Surprisingly, we see that the best performing model is S3, that was fine-tuned only for 1 epoch. This means that the model is showing overfitting signs in a very early despite the high number of training examples. A comparison to recent literature, in particular to the 2021 MEDIQA winner, described on Abacha et al. [1], shows that although we do achieve a performance we are still distant from the state-of-the-art result.

Table 6. Results - Summarization Task ROUGE-1

	ROUGE – 1		
	Precision	*Recall*	*F1 – Score*
Base	0.19412	0.25569	0.19623
*S*1	0.51951	0.38811	0.41816
*S*2	0.51781	0.40390	0.42788
*S*3	0.52231	0.39212	0.42263
L[1]	-	-	0.55730

Table 7. Results - Summarization Task ROUGE-2

	ROUGE – 2		
	Precision	*Recall*	*F1 – Score*
Base	0.08357	0.09791	0.07938
*S*1	0.34269	0.25308	0.27133
*S*2	0.33146	0.25800	0.27081
*S*3	0.35009	0.2611	0.27947
L[1]	-	-	0.43620

Table 8. Results - Summarization Task ROUGE-L

	ROUGE – *L*		
	Precision	*Recall*	*F1 – Score*
Base	0.16932	0.23041	0.17435
*S*1	0.49906	0.37408	0.40278
*S*2	0.49644	0.38779	0.41069
*S*3	0.50317	0.37778	0.40746
L[1]	-	-	0.53660

6. Conclusions

Our work had the purpose of exploring the potential of new and powerful model architectures ,like the Text-to-Text Transfer Transformer, when applied to the medical domain. We developed a multi-task medical model capable of dealing with 4 different tasks: summarization of radiology reports, medical semantic text similarity, medical natural language inference, and medical paraphrase detection. We also presented a training scheme that takes advantage of the differences in the nature of the datasets sizes and optimizes the time required to train a model. Our fine-tuning approach revealed significant results on the mentioned tasks, similar to state-of-the-art models, but with the ability to generalize for 4 different tasks. We also demonstrated the potential of data augmentation for the medical domain by applying a back translation strategy and by generating new examples based on the radiology report labelling task outputs of the ChexBERT Model. Our models suffered from one of the most common problems during training of medical models, overfitting of the training data, due to the problem of scarce datasets that were available. The results achieved with our last model, show significant improvement regarding all the three small tasks and summarization task. In particular, according to the official results table from the 2021 MEDIQA task, our model would be rated on the 10th place, without any modification to the model architecture, which indicates the robustness of our training scheme and potential for improvements which take advantage purely from training strategies and data augmentation.

7. Future Work

The application of models to the medical domain is still in a very early stage. Future work to improve models like T5 still needs to be done. Much of the research work is spent researching new architectures, which shifts the focus of experimenting and achieving the best results with the current ones. Efforts should pass by exploring simple ideas, like exploring different training strategies such as on how to take full advantage of multi-task learning. A possible modification in our models to be explore is to add different training objectives related with the tasks to explore. For example, instead of minimizing the loss for summarization, to modify this objective to minimize a combination of loss and a metric such as ROUGUE or QAEval [5]. Due to restrictions of resources, our experiments on the decoding phase were incomplete and are not reported on this document. This particular step also showed great promising. Future work should explore the the use of re-ranking techniques to select the best output based on a metric of our interest, This allows us to explore different outputs of our models. Techniques like beam search and sampling with multiple returned sentences allow for a multitude methods that can applied as an aide to the inference of our model. Another relevant topic we manage to explore fully and it is not described here is conditional decoding, where we restrict the type of tokens the model was able to predict, taken into account the task. This strategy seemed particularly promising for tasks like STS, were a very short output was generated. Finally, we suggest further exploration on a hot topic in the NLP community nowadays, which is quality evaluation. This paradigm is particularly relevant in machine translation and could potentially by applied to generation as a way of measuring good or bad summaries.

Acknowledgement

I would like to thank professor Bruno Martins and Dr. Nuno André da Silva, for the opportunity to explore this topic. In particular, I would like to thank Prof. Bruno for the incredible patience and dedication to supervise this work closely.

References

[1] Asma Ben Abacha, Yassine M'rabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. Overview of the mediqa 2021 shared task on summarization in the medical domain. pages 74–85, 6 2021. doi: 10.18653/V1/2021.BIONLP-1.8. URL https://aclanthology.org/2021.bionlp-1.8.

[2] A Aditya. Paraphrase detection using deep learning. 2018.

[3] Rahul Bhagat and Eduard Hovy. What is a paraphrase? *Computational Linguistics*, 39:463–472, 9 2013. ISSN 0891-2017. doi: 10.1162/COLI_A_00166. URL http://direct.mit.edu/coli/article-pdf/39/3/463/1801912/coli_a_00166.pdf.

[4] Jean-Philippe Corbeil, Polytechnique Montreal, and Hadi Abdi Ghadivel. Bet: A backtranslation approach for easy data augmentation in

transformer-based paraphrase identification context a preprint. 2020. URL https://www.kaggle.com/c/quora-question-pairs.

[5] Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. Towards question-answering as an automatic metric for evaluating the content quality of a summary, 2021.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[7] William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL https://aclanthology.org/I05-5002.

[8] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pages 590–597, 1 2019. URL https://arxiv.org/abs/1901.07031v1.

[9] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Mahdi Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3, 2016.

[10] Alistair E.W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6, 12 2019. ISSN 20524463. doi: 10.1038/s41597-019-0322-0.

[11] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, July 2020. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

[12] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

[13] Xing Meng, Craig H Ganoe, Ryan T Sieberg, Yvonne Y Cheung, and Saeed Hassanpour. Self-supervised contextual language representation of radiology reports to improve the identification of communication urgency. 2019.

[14] Long N Phan, James T Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadıro ˘ Glu, Alec Peltekian, and Grégoire Altan-Bonnet. Scifive: a text-to-text transformer model for biomedical literature. 2021. URL https://www.ncbi.nlm.nih.gov/pmc.

[15] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. 10 2019. URL http://arxiv.org/abs/1910.10683.

[16] Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 1586–1596, 2020.

[17] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. URL https://arxiv.org/pdf/2004.09167.pdf.

[18] Amber Stubbs, Michele Filannino, Ergin Soysal, Samuel Henry, and Özlem Uzuner. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *Journal of the American Medical Informatics Association*, 26 (11):1163–1171, 09 2019. ISSN 1527-974X. doi: 10.1093/jamia/ocz163. URL https://doi.org/10.1093/jamia/ocz163.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 6 2017. URL http://arxiv.org/abs/1706.03762.

[20] Yuxia Wang, Karin Verspoor, and Timothy Baldwin. Learning from unlabelled data for clinical semantic textual similarity. *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 227–233, November 2020. doi: 10.18653/v1/2020.clinicalnlp-1.25. URL https://aclanthology.org/2020.clinicalnlp-1.25.

[21] Xi Yang, Xing He, Hansi Zhang, Yinghan Ma, Jiang Bian, and Yonghui Wu. Measurement of semantic textual similarity in clinical texts: Comparison of transformer-based models. 202. doi: 10.2196/19735. URL http://medinform.jmir.org/2020/11/e19735/.

[22] Yanshan, Sunyang Fu, Feichen Shen, Sam Henry, Ozlem Uzuner, and Hongfang Liu. The 2019 n2c2/ohnlp track on clinical semantic textual similarity: Overview. *JMIR Med Inform 2020;8(11):e23375 https://medinform.jmir.org/2020/11/e23375*, 8:e23375, 11 2020. doi: 10.2196/23375. URL https://medinform.jmir.org/2020/11/e23375.

[23] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. 2020. URL https://arxiv.org/pdf/1912.08777.pdf.