Cross-Domain Transfer of Generative Explanations using Text-to-Text Models

Karl Fredrik Erliksson[1,2], Anders Arpteg[2],
Mihhail Matskin[1], and Amir H. Payberah[1]

[1] KTH Royal Institute of Technology, Sweden
kferl@kth.se, misha@kth.se, payberah@kth.se
[2] Peltarion, Sweden
anders@peltarion.com

Abstract. Deep learning models based on the Transformers architecture have achieved impressive state-of-the-art results and even surpassed human-level performance across various natural language processing tasks. However, these models remain opaque and hard to explain due to their vast complexity and size. This limits adoption in highly-regulated domains like medicine and finance, and often there is a lack of trust from non-expert end-users. In this paper, we show that by teaching a model to generate explanations alongside its predictions on a large annotated dataset, we can transfer this capability to a low-resource task in another domain. Our proposed three-step training procedure improves explanation quality by up to 7% and avoids sacrificing classification performance on the downstream task, while at the same time reducing the need for human annotations.

Keywords: Explainable AI · Generative explanations · Transfer learning

1 Introduction

There is a growing consensus that many practical machine learning (ML) applications require explainability, especially when these applications are subject to critical auxiliary criteria that are difficult to formulate mathematically, e.g., nondiscrimination, safety, or fairness [11, 30]. Moreover, regulations such as the General Data Protection Regulation (GDPR) [13] equip people with a "right to explanation" for algorithmic decisions that significantly affect them. At the same time, deep neural networks (NNs) have achieved and even surpassed human performance in many tasks in natural language processing (NLP) and computer vision [15, 43], which has motivated a large body of research over the last few years focusing on making NN predictions more explainable.

Explainability in ML has traditionally been approached from two perspectives; either by building models that provide inherent transparency and explainability [5, 21, 26], or by creating post-hoc explanations for an opaque model that has already been trained [29, 37, 39]. This work falls into the former category where we teach a model to generate explanations as part of the prediction process, conceptually similar to how humans would be asked to motivate

their reasoning for a specific decision. The explanations are formed by natural language, and we cast this as a supervised sequence-to-sequence (seq2seq) problem where the model learns from ground-truth explanations annotated by humans [4, 36, 40]. Natural language explanations provide a series of benefits compared to other common approaches, such as attributions methods and formal language. They are more easily accessible to non-expert end-users owing to the familiar format [4], and are often simpler to evaluate and annotate by humans. Narang et al. [32] recently investigated this approach and proposed a model called WT5 that achieves new state-of-the-art performance on various NLP explainability benchmarks [10]. However, this requires large amounts of annotated explanations during training and for many real-world applications this becomes a bottleneck.

We propose a three-step training procedure to transfer the ability to generate extractive explanations from a large easily-available dataset to a low-resource downstream task with a lack of annotated ground-truth explanations, in a potentially different domain. First, in the pre-training (PT) step, we train an initial language model using unannotated data. Then, in the explainability pre-training (EP) step, we teach the model the semantic meaning of an *explainability keyword*. Finally, we use this keyword during the fine-tuning (FT) step and at inference time to instruct the model to generate explanations for specific predictions. To summarize our contributions:

- Narang et al. in [32] provide a brief qualitative discussion regarding explainability transfer for WT5. We extend this work and provide a more thorough quantitative evaluation, including two popular seq2seq models, T5 [35] and BART [27]. We find that T5 consistently outperforms BART for extractive explanation generation across all our experiments.
- Using our proposed three-step training procedure, we show that the ability to generate extractive explanations can be transferred between tasks in different domains, and that it can result in both improved performance and explanation quality on a low-resource downstream task with few annotated explanations.
- We provide evidence that only a small number of samples from the downstream tasks need to be annotated with human explanations to achieve a significant boost in explanation quality.

Through the experiments, we see an increase of 7% and 5% in TF1 score (explanation quality) for T5-Base and T5-Large, respectively, when EP is performed.[3]

2 Background

In this section, we provide a brief background to seq2seq modelling in NLP and define the main idea of generative explanations.

2.1 Sequence-to-Sequence Models

Consider an NLP model $f : \mathcal{X} \to \mathcal{Y}$ where the input $x = (x_1, x_2, ..., x_{N_{\text{in}}}) \in \mathcal{X}$ and the output $y = (y_1, y_2, ..., y_{N_{\text{out}}}) \in \mathcal{Y}$ are both ordered sequences of tokens.

[3] Code available at https://github.com/Peltarion/explainability_transfer

By $\tilde{x}$ and $\tilde{y}$, we denote the corresponding raw input and output text. The model f is trained by maximizing the the conditional probability $p(y_1, ..., y_{N_{\text{out}}}|x_1, ..., x_{N_{\text{in}}}) = \prod_i^{N_{\text{out}}} p(y_i|x_1, ..., x_{N_{\text{in}}}, y_1, ..., y_{i-1})$. At prediction time, an output sequence can be generated autoregressively by iteratively sampling $y_i \sim p(y_i|x_1, ..., x_{N_{\text{in}}}, y_1, ..., y_{i-1})$ either greedily or by methods like beam search.

Raffel et al. [35] introduced the idea of unifying all NLP tasks into a general common framework by treating them as seq2seq problems, referred to as the *text-to-text* framework. As an example, a binary classification problem with output classes {`True`, `False`} is posed as a generative task where the model is trained to explicitly generate the sequence of tokens corresponding to the target output class. This should be seen in contrast to other common BERT-based architectures [9], where a small model head tailored for a specific task and its format is attached on top of an encoder block to produce a probability distribution over the output classes. The raw input is formatted as $\tilde{x}$ = "⟨`task_prefix`⟩: ⟨`input_text`⟩", where the prefix is used to let the model know what type of task it is, e.g., "`sentiment`" for sentiment analysis. The target output is given by $\tilde{y}$ = "⟨`target`⟩", which in the case of classification problems would simply be the class label. This enables an easy way of transferring knowledge from one task to the other, thanks to the unified format. If the model would output anything other than the expected output classes during evaluation, it is considered as incorrect.

The Text-to-Text Transfer Transformer (T5) [35] is a model based on the above approach that was pre-trained on the large Common Crawl dataset [7], and has been demonstrated to achieve state-of-the-art performance on various NLP downstream tasks [43]. Apart from T5, many other seq2seq models have been used for tasks such as machine translation and text summarization. A recent popular model is BART [27], which is architecturally similar to T5 but using a different language model pre-training objective and number of hidden states in the embedding and feed-forward layers.

2.2 Generative Explanations

One way to approach explainability in deep learning is by letting a model produce explanations similar to how humans would motivate their reasoning. One of the earlier works by Hendricks et al. [16] considered generating "`because of`" sentences for a computer vision classification task. The text-to-text framework enables a new way to teach NLP models to produce generative explanations in a supervised fashion. This idea was recently explored in [32], where an extension of T5, called WT5 (short for "Why T5?"), was proposed. In this case, we simply prepend ⟨`task_prefix`⟩ in $\tilde{x}$ with the optional keyword "`explain`" and append the target output $\tilde{y}$ with "`explanation`: ⟨`explanation`⟩", where we assume that golden-truth annotated explanations are available for the task. The new input-output format thus becomes

$$\begin{aligned} \tilde{x} &= \text{“[}\texttt{explain}\text{] ⟨}\texttt{task\_prefix}\text{⟩: ⟨}\texttt{input\_text}\text{⟩”}, \\ \tilde{y} &= \text{“⟨}\texttt{target}\text{⟩ [}\texttt{explanation}\text{: ⟨}\texttt{explanation}_1\text{⟩)] ...} \\ &\quad \text{[}\texttt{explanation}\text{: ⟨}\texttt{explanation}_M\text{⟩)]”}, \end{aligned} \tag{1}$$

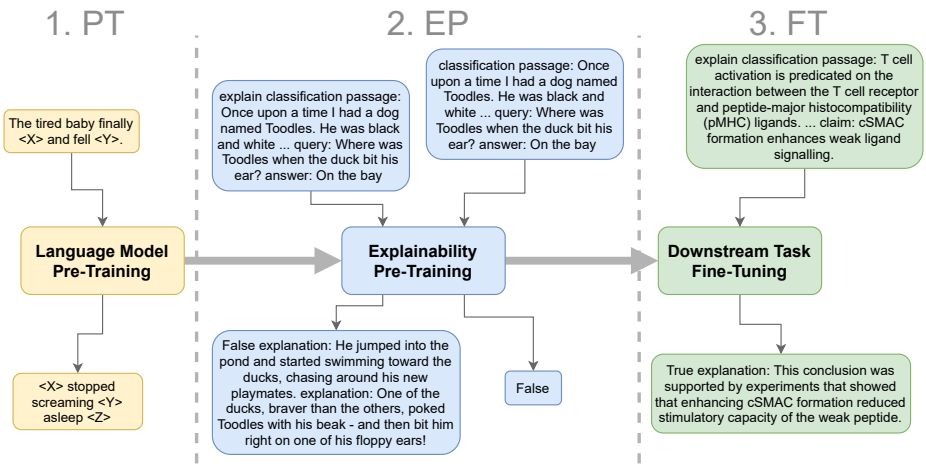


Fig. 1: The proposed three-step training procedure.

where hard brackets denote optional explanation arguments and we allow for potentially multiple explanation sentences. An illustrative example of the input-output format is provided in Table 1. To simplify the annotation and evaluation process, it is helpful to consider the subset of *extractive explanations* that only consist of spans of tokens from the input text. This allows us to compute overlap statistics with respect to the ground truth to quantitatively measure the explanation quality [10].

3 Approach

The main focus of this work is to transfer explainability capabilities to a low-resource task in another domain with a potentially limited number of annotated explanations. Based on the procedure outlined in [32], we utilize seq2seq models to generatively produce natural language explanations alongside the original prediction task. To this end, we propose a three-step training procedure as illustrated in Figure 1:

1. *Language model pre-training (PT)* is carried out in a self-supervised fashion on a large text corpus like C4 [35] (the yellow blocks in Figure 1).[4]
2. *Explainability pre-training (EP)* is then performed on a large dataset with annotated explanations (the blue blocks in Figure 1). Following the ideas in [32], we teach the model the meaning of the "`explain`" keyword by uniformly at random constructing training instances with and without annotated explanations according to the format in equation (1). We hypothesize that this promotes a task-agnostic extractive explanation capability that can be extended also for various other tasks.
3. *Fine-tuning (FT) on the downstream task* is carried out with as many annotated explanations as are available (the green blocks in Figure 1). At

[4]Since all seq2seq models considered in this work have publicly released checkpoints from language model pre-training, this is used as starting point for step 2 in Figure 1.

prediction time and during evaluation, the "`explain`" keyword is prepended to all instances, thus instructing the model to always generate explanations alongside its predictions.

Conceptually, there are no specific assumptions on the domain or semantics of the FT task, thus allowing the framework to be applicable broadly. To facilitate transferability, we consider FT tasks that can be cast into a similar input-output format as during the EP step, in this work text-classification problems.

4 Experiments

In this section, we first introduce the datasets, tasks, and evaluation metrics, and then evaluate our proposed approach for transferring generative explanation capabilities between tasks in potentially different domains. We do this in two different settings: (1) with all available annotated explanations, and (2) with limited annotated explanations during FT.

4.1 Datasets

We use three datasets in our experiments:

1. MultiRC [20][5]: a reading comprehension dataset consisting of multiple-choice questions for short paragraphs of text with annotated supporting evidence spans. We consider the binary classification of a given question and answer candidate pair.
2. FEVER [40][5]: a large fact verification dataset extracted from Wikipedia that has been annotated by humans with supporting evidence spans. We consider claims that are either *supported* or *refuted*.
3. SciFact [42]: a small dataset where the task is to find abstracts from a corpus of research literature, and corresponding evidence sentences, that *support* or *refute* scientific and medical claims. We consider the subtask of text classification for a given claim-abstract pair and use the corresponding evidence sentences as ground-truth extractive explanations. Abstracts that do not contain any evidence for a claim are discarded, making the classification problem binary.

We use MultiRC and FEVER during the EP step and SciFact as the final downstream FT task, thus considering transfer from general English to the scientific and medical domain. To unify the input-output format and simplify transferability, all tasks are cast as binary classification problems where the output labels are {`True`, `False`}.

4.2 Evaluation

Consider the generic target output format for any of the introduced tasks,

$$\tilde{y} = y_{\text{label}} \;\; \texttt{explanation:}\; e_1 \; ... \; \texttt{explanation:}\; e_M, \tag{2}$$

where y_{label} is the target label, either `True` or `False`, and $\mathcal{E} = \{e_1, ..., e_M\}$ is the ground-truth explanation consisting of M sentences. The predicted output sequence $\hat{y}$ is assumed to follow the desired format and is split by the

[5]We use the dataset versions distributed through the ERASER benchmark [10].

Table 1: Illustrative example of data post-processing and explanation quality evaluation metrics. Overlap spans are highlighted in gray.

Var.	Value
$\tilde{x}$	"explain classification passage: I had a dog named Toodles. He was black and white and had long floppy ears. He also had very short legs. Every Saturday we would go to the park and play Toodles' favorite game. query: What describes Toodles' legs? answer: Long"
$\tilde{y}$	"False explanation: I had a dog named Toodles. explanation: He also had very short legs."
$\hat{y}$	"False explanation: I had a dog called Toodles. explanation: He also had very short legs."
$\mathcal{E}$	{"I had a dog named Toodles.", "He also had very short legs."}
$\hat{\mathcal{E}}$	{"I had a dog called Toodles.", "He also had very short legs."}
$\mathcal{S}$	"explain classification passage: I had a dog named Toodles. He was black and white and had long floppy ears. He also had very short legs. Every Saturday we would go to the park and play Toodles' favorite game. query: What describes Toodles' legs? answer: Long"
$\hat{\mathcal{S}}$	"explain classification passage: I had a dog named Toodles. He was black and white and had long floppy ears. He also had very short legs. Every Saturday we would go to the park and play Toodles' favorite game. query: What describes Toodles' legs? answer: Long"
P: 100.00%	**R**: 50.00% **TF1**: 66.67% **BLEU**: 84.92% **ROUGE-L**: 93.33%

"`explanation:`" separator to form the predicted label $\hat{y}_{\text{label}}$ and explanation set $\hat{\mathcal{E}} = \{\hat{e}_1, ..., \hat{e}_{\hat{M}}\}$. If the model would output anything other than the desired format, this would be counted as part of the predicted label and thus resulting in both poor task performance and explanation quality.

We use four evaluation metrics in our experiments: F1 score for prediction task performance, as well as token-level F1 score (TF1), BLEU score [33], and ROUGE-L score [28] to measure extractive explanation quality. Each explanation sentence $e \in \mathcal{E}$ is tokenized and matched against all possible spans in the tokenized input text $\tilde{x}$. This forms a corresponding set of overlap tuples $\mathcal{S} = \{(e_{i_{\text{start}}}, e_{i_{\text{end}}}) \mid e \in \mathcal{E}\}$ of the start and end indices of the matched spans, and analogously $\hat{\mathcal{S}}$ from $\hat{\mathcal{E}}$. If an explanation does not exactly match any span, it is considered invalid and is discarded. TF1 is computed as the F1 score between $\hat{\mathcal{S}}$ and $\mathcal{S}$, averaged over all N samples in the dataset:

$$\text{TF1} = \frac{1}{N}\sum_{k=1}^{N}\frac{\text{P}_k \cdot \text{R}_k}{\text{P}_k + \text{R}_k}, \quad \text{P}_k = \frac{|\mathcal{S}^{(k)} \cap \hat{\mathcal{S}}^{(k)}|}{|\hat{\mathcal{S}}^{(k)}|}, \quad \text{R}_k = \frac{|\mathcal{S}^{(k)} \cap \hat{\mathcal{S}}^{(k)}|}{|\mathcal{S}^{(k)}|}. \tag{3}$$

The TF1 score significantly punishes generated outputs that deviate from the desired format, or if a generated explanation sentence is not exactly matching a span in the input text. To make the evaluation more nuanced, we also compute BLEU score and ROUGE-L score directly between the raw output text $\hat{y}$ and $\tilde{y}$. These metrics measure precision and recall-based overlap statistics, respectively, between shorter spans of different lengths and are not as binary as TF1. BLEU score has been previously used for abstractive explanation evaluation [4, 32]. An illustrative example of the data post-processing procedure and the evaluation metrics are provided in Table 1.

Random Baseline. To put our results into a quantitative context, we construct a random baseline for each task. This is achieved by randomly sampling a predicted label according to the class weights in the training dataset. Additionally,

we empirically estimate the probability mass function of the number of sentences M that constitute the extractive explanations in the training dataset. To form $\hat{\mathcal{E}}$, $\hat{M}$ is sampled independently from this distribution for each instance in the evaluation dataset, and the corresponding number of explanation sentences are then selected uniformly at random from the input text.

4.3 Model and Training Details

We consider two seq2seq models based on the Transformers architecture [41], namely T5 [35] and BART [27]. We analyze both the Base and the Large variants of T5 and the Large variant of BART. The experimental setup follows the training procedure outlined in Figure 1, where MultiRC and FEVER are used during EP and SciFact is the FT task.

To teach the model to explain its predictions, EP instances are sampled with equal probability from a mixture of training samples with and without annotated explanations. Every time an explanation is added to the target output, the input text is prepended with the "`explain`" keyword as described in Section 2. This allows the model to learn the semantic meaning of the "`explain`" keyword, and the same format can be used during FT to generate explanations. We evaluate the model every 360 steps on the evaluation dataset and the checkpoint that achieves the lowest F1 score is used for further fine-tuning on the downstream task. After fine-tuning, average F1 and TF1 score is used as the final evaluation metric to select the best model checkpoint. For T5-Base and BART-Large, we repeat all experiments five times, and for T5-Large three times due to its large size and needed computational effort.[6]

Table 2: Validation set performance on SciFact with all annotated explanations.

Model	EP	F1	TF1	BLEU	ROUGE-L
T5-Large	None	84.0 (±2.9)	66.4 (±1.7)	71.9 (±1.3)	77.4 (±0.9)
	MultiRC	86.7 (±2.1)	**69.4** (±1.4)	73.2 (±1.4)	78.3 (±1.5)
	FEVER	88.4 (±1.6)	69.0 (±1.0)	74.3 (±2.2)	**79.2** (±0.5)
T5-Base	None	78.5 (±0.4)	64.6 (±0.7)	71.3 (±1.3)	75.8 (±0.5)
	MultiRC	81.9 (±1.4)	68.2 (±1.8)	**74.9** (±2.2)	78.4 (±2.1)
	FEVER	85.3 (±0.9)	69.2 (±0.5)	74.3 (±0.6)	78.8 (±0.3)
BART-Large	None	61.0 (±4.0)	37.7 (±6.6)	40.7 (±6.3)	57.7 (±7.0)
	MultiRC	85.8 (±2.2)	46.2 (±1.2)	42.9 (±1.8)	65.8 (±2.5)
	FEVER	**90.0** (±1.5)	45.0 (±0.6)	40.0 (±5.2)	64.5 (±3.5)
Random baseline	None	67.5 (±2.7)	19.1 (±1.8)	25.5 (±2.0)	32.4 (±1.7)

4.4 All Available Annotated Explanations for SciFact

Table 2 shows the results after FT with all available annotated explanations for SciFact. As a quantitative reference, we include a baseline for each model type when EP is not performed. These results are not directly comparable with [42], since we consider the subtask of label prediction and rationalization for the subset of refuted and supported claims. Overall, the T5-based models achieve

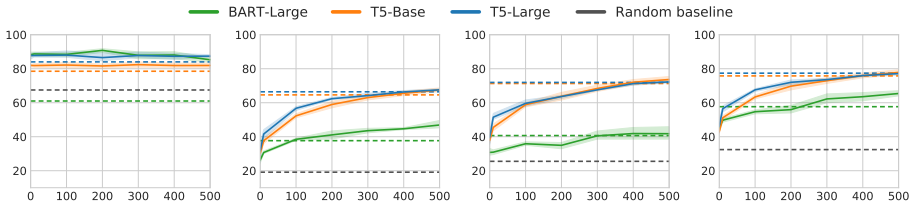
[6] The hyperparameter settings for the different models and training phases are available in the public code repository.

Table 3: Non-cherry picked samples from the SciFact validation set for WT5-Large after MultiRC EP. Explanations in $\mathcal{S} \cap \hat{\mathcal{S}}$ are highlighted in green, $\hat{\mathcal{S}} \setminus \mathcal{S}$ in yellow, and $\mathcal{S} \setminus \hat{\mathcal{S}}$ in red (not present). The remaining part of the input text has been shortened.

Claim	**Prediction**
Taxation of sugar-sweetened beverages had no effect on the incidence rate of type II diabetes in India	**False**
BACKGROUND Taxing sugar-sweetened beverages (SSBs) has been proposed in high-income countries to reduce obesity and type 2 diabetes. ... **The 20% SSB tax was anticipated to reduce overweight and obesity prevalence by 3.0% (95% CI 1.6%-5.9%) and type 2 diabetes incidence by 1.6% (95% CI 1.2%-1.9%) among various Indian subpopulations over the period 2014-2023, if SSB consumption continued to increase linearly in accordance with secular trends. However, acceleration in SSB consumption trends consistent with industry marketing models would be expected to increase the impact efficacy of taxation, averting 4.2% of prevalent overweight/obesity (95% CI 2.5-10.0%) and 2.5% (95% CI 1.0-2.8%) of incident type 2 diabetes from 2014-2023.** ... **CONCLUSION Sustained SSB taxation at a high tax rate could mitigate rising obesity and type 2 diabetes in India among both urban and rural subpopulations.**	
Macrolides have no protective effect against myocardial infarction	**True**
CONTEXT Increasing evidence supports the hypothesis of a causal association between certain bacterial infections and increased risk of developing acute myocardial infarction. ... **No effect was found for previous use of macrolides (primarily erythromycin), sulfonamides, penicillins, or cephalosporins.** ...	
Stroke patients with prior use of direct oral anticoagulants have a lower risk of in-hospitality mortality than stroke patients with prior use of warfarin	**False**
Importance Although non-vitamin K antagonist oral anticoagulants (NOACs) are increasingly used to prevent thromboembolic disease, there are limited data on NOAC-related intracerebral hemorrhage (ICH). ... **The unadjusted in-hospital mortality rates were 32.6% for warfarin, 26.5% for NOACs, and 22.5% for no OACs. Compared with patients without prior use of OACs, the risk of in-hospital mortality was higher among patients with prior use of warfarin (adjusted risk difference [ARD], 9.0% [97.5% CI, 7.9% to 10.1%]; adjusted odds ratio [AOR], 1.62 [97.5% CI, 1.53 to 1.71]) and higher among patients with prior use of NOACs (ARD, 3.3% [97.5% CI, 1.7% to 4.8%]; AOR, 1.21 [97.5% CI, 1.11-1.32]). Compared with patients with prior use of warfarin, patients with prior use of NOACs had a lower risk of in-hospital mortality (ARD, -5.7% [97.5% CI, -7.3% to -4.2%]; AOR, 0.75 [97.5% CI, 0.69 to 0.81]).** ... **Prior use of NOACs, compared with prior use of warfarin, was associated with lower risk of in-hospital mortality.**	

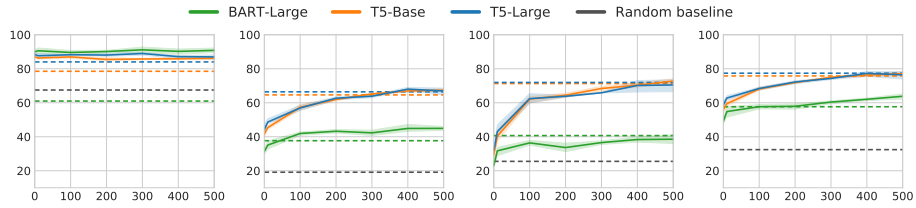
significantly higher explanation quality compared to BART-Large, and we see consistent performance gains across all metrics when MultiRC or FEVER are used for EP. T5-Large achieves the highest TF1 score with a relative gain of 5%, closely followed by T5-Base that sees a relative gain of 7%. EP using FEVER has the highest positive impact on the prediction task performance (F1 score).

To provide a qualitative understanding of the generated explanations, three non-cherry picked examples for T5-Large with MultiRC during EP are shown in Table 3. The first claim is correctly refuted and the model generates all three sentences of the golden annotated explanation. The second claim is also correctly classified and in this case only one sentence constitutes both the predicted and golden explanation, which illustrates the flexibility in the generative approach. The last example is classified incorrectly, even though the model is extracting a majority of the actual golden explanation. This suggests two possible reasons; that the model is able to find the relevant part of the input but cannot infer the correct label from this, or that it generates a plausible explanation even though this is actually not used in the label-prediction process. Since the training loss



(a) F1 (b) TF1 (c) BLEU (d) ROUGE-L

Fig. 2: Explainability transfer from MultiRC to SciFact. Evaluation metrics (a)–(d) with 95% confidence intervals as a function of number of annotated explanations during FT. Dashed lines correspond to the same values as EP *None* in Table 2.



(a) F1 (b) TF1 (c) BLEU (d) ROUGE-L

Fig. 3: Explainability transfer from FEVER to SciFact.

function encourages the same extractive explanations regardless of the label, there are no theoretical guarantees for explanation faithfulness. Wiegreffe et al. [45] investigate this phenomenon and provides some empirical evidence that there is indeed a robustness between the generated explanations and labels, but that further work in this area is needed.

4.5 Downstream Task with Limited Annotated Explanations

For most practical applications, annotated explanations on the target downstream task are scarce and costly to obtain. To evaluate the effectiveness of explainability transfer to alleviate these problems, we simulate scenarios with different number of available annotated explanations on SciFact. In all cases, EP is performed using all available explanations. Figure 2 depicts transfer from MultiRC to SciFact using $n_{\texttt{exp}} \in \{0, 10, 100, 200, 300, 400, 500\}$ out of 546 annotated training samples for SciFact.

For all models, there is an increase in prediction performance (F1 score) of performing EP, and it stays more or less constant regardless of $n_{\texttt{exp}}$. For the T5 models, we also see improved explanation quality across all metrics. This suggests that the EP procedure allows the model to be fine-tuned more effectively so that the WT5 explanation framework does not sacrifice task performance. As the number of annotated explanations approach zero, the explanation quality drops drastically, which indicates that zero-shot explainability transfer is indeed challenging. However, with just 200 annotated samples corresponding to roughly 35% of the training dataset, T5-Large achieves strong explanation quality almost matching the baseline with all available annotated explanations. Generally, T5-

Base achieves nearly identical explanation quality metrics as T5-Large, however with slightly worse prediction task performance. This is surprising since T5-Large achieved higher explanation quality during EP on both MultiRC and FEVER. We believe that the small size of SciFact might benefit the smaller base model to more effectively transfer the explanation capability to the new task.

The explanation quality for BART is considerably lower than the T5 counterparts, meanwhile the prediction task performance is still competitive. BART is not as good at conforming to the strict extractive explanation format, which hurts the TF1 and BLEU score. ROUGE-L is also inferior but the gap to the T5 models is not as significant. We provide corresponding results for explainability transfer from FEVER to SciFact in Figure 3, which follow the same general trends.

5 Related Work

Explainable ML has received a lot of research interest over the last few years and a comprehensive review of the field in general is provided in [14] and specifically in [8] related to applications for NLP. This work belongs to a class of methods that provide explainability by design and more specifically self-explaining systems, where the model itself produces an explanation as part of the prediction process. Attention-based models have mainly been considered for this purpose in NLP [6, 23, 44], much owing to the recent success of the Transformers architecture and the hope that this offers some inherent explainability "for free". However, the usefulness and validity of attention weights as explanations have been questioned [3, 19, 38].

Generative natural language explanations were studied in [4], who proposed an extended version of the SNLI dataset [2] with annotated abstractive explanations, and considered different seq2seq models for learning to generate such explanations. This work is based on [32], which approached the same problem by casting it into the T5 text-to-text framework [35]. Other previous works have also studied generative explanations for non-NLP tasks [12, 16, 22].

Another line of work for explainable NLP is based on rationalization pipelines that aim to produce extractive explanations by splitting the prediction process into two subsequent modules; a rationale extractor and a predictor [1, 10, 25, 34]. The benefit of this approach is that it provides some faithfulness guarantees by construction since the predictor can only rely on the extracted rationales, however, potentially at the expense of prediction performance. The dilemma of faithful and plausible explanations was raised in [17] and was further studied in [45] for generative explanations. Both argue that self-explaining systems, although not guaranteedly faithful, can still be very useful in practice.

In the medical domain specifically, 1-dimensional CNNs with label-conditional attention have been explored for explainable ICD code prediction from discharge summaries [31]. The interpretability of Transformer attention weights in a medical context was analyzed and questioned in [18]. Recently, rationalization pipelines have been applied to medical and scientific text, for instance [42] and [24] utilize BERT-to-BERT models for SciFact and for classifying random clinical trials, respectively.

6 Conclusions

In this work, we have demonstrated that generating extractive explanations can be transferred from general English to tasks in the scientific and medical domain. Our proposed three-step training procedure with explainability pre-training improves explanation quality as well as prediction task performance on the downstream task. Furthermore, we see a large increase in explanation quality for only a small number of annotated explanations during fine-tuning, making it an attractive option for real-world use cases where annotations are limited and costly to obtain. An interesting direction for future work is to analyze the impact of specific weights of the classification and explanation objectives in the common loss function. We plan to shed further light on the faithfulness-plausibility dilemma by applying attribution methods (e.g., SHAP [29]) on top of the generated explanations. The practical usability of the generated explanations will also be further assessed by human evaluation studies. As an extension to cross-domain explainability transfer, the same approach can also be considered for explainability transfer across languages. We believe recent multilingual seq2seq models like mT5 [46] to be a promising candidate for this purpose.

References

1. Bastings, J., et al.: Interpretable neural predictions with differentiable binary variables. In: ACL (2019)
2. Bowman, S.R., et al.: A large annotated corpus for learning natural language inference. In: EMNLP (2015)
3. Brunner, G., et al.: On identifiability in transformers. In: ICLR (2019)
4. Camburu, O., et al.: e-snli: Natural language inference with natural language explanations. In: NeurIPS (2018)
5. Chen, C., et al.: This looks like that: Deep learning for interpretable image recognition. In: NeurIPS (2019)
6. Clark, K., et al.: What does BERT look at? An analysis of bert's attention. In: ACL BlackboxNLP Workshop (2019)
7. Common Crawl: https://www.commoncrawl.org
8. Danilevsky, M., et al.: A survey of the state of explainable AI for natural language processing. In: AACL-IJCNLP (2020)
9. Devlin, J., et al.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL (2019)
10. DeYoung, J., et al.: ERASER: A benchmark to evaluate rationalized NLP models. In: ACL (2020)
11. Doshi-Velez, F., et al.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
12. Ehsan, U., et al.: Rationalization: A neural machine translation approach to generating natural language explanations. In: AIES (2018)
13. EU: General Data Prodection Regulation (GDPR): Recital 71 (2018), https://www.privacy-regulation.eu/en/r71.htm
14. Guidotti, R., et al.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5) (2018)
15. He, K., et al.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: ICCV (2015)

16. Hendricks, L., et al.: Generating visual explanations. In: ECCV (2016)
17. Jacovi, A., et al.: Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In: ACL (2020)
18. Jain, S., et al.: An analysis of attention over clinical notes for predictive tasks. In: ClinicalNLP (2019)
19. Jain, S., et al.: Attention is not explanation. In: NAACL (2019)
20. Khashabi, D., et al.: Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In: NAACL (2018)
21. Kim, B., et al.: The bayesian case model: A generative approach for case-based reasoning and prototype classification. In: NIPS (2014)
22. Kim, J., et al.: Textual explanations for self-driving vehicles. In: ECCV (2018)
23. Kovaleva, O., et al.: Revealing the dark secrets of BERT. In: NeurIPS (2019)
24. Lehman, E., et al.: Inferring which medical treatments work from reports of clinical trials. In: NAACL (2019)
25. Lei, T., et al.: Rationalizing neural predictions. In: EMNLP (2016)
26. Letham, B., et al.: Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. Annals of Applied Statistics **9** (2015)
27. Lewis, M., et al.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: ACL (2020)
28. Lin, C.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out (2004)
29. Lundberg, S., et al.: A unified approach to interpreting model predictions. In: NIPS (2017)
30. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267** (2019)
31. Mullenbach, J., et al.: Explainable prediction of medical codes from clinical text. In: NAACL (2018)
32. Narang, S., et al.: WT5?! Training text-to-text models to explain their predictions. arXiv preprint arXiv:2004.14546 (2020)
33. Papineni, K., et al.: BLEU: A method for automatic evaluation of machine translation. In: ACL (2002)
34. Paranjape, B., et al.: An information bottleneck approach for controlling conciseness in rationale extraction. In: EMNLP (2020)
35. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR (2020)
36. Rajani, N., et al.: Explain yourself! Leveraging language models for commonsense reasoning. In: ACL (2019)
37. Ribeiro, M., et al.: "Why should i trust you?" Explaining the predictions of any classifier. In: KDD (2016)
38. Serrano, S., et al.: Is attention interpretable? In: ACL (2019)
39. Sundararajan, M.: Axiomatic attribution for deep networks. In: ICML (2017)
40. Thorne, J., et al.: FEVER: A large-scale dataset for fact extraction and verification. In: NAACL (2018)
41. Vaswani, A., et al.: Attention is all you need. NIPS (2017)
42. Wadden, D., et al.: Fact or fiction: Verifying scientific claims. In: EMNLP (2020)
43. Wang, A., et al.: Superglue: A stickier benchmark for general-purpose language understanding systems. In: NeurIPS (2019)
44. Wiegreffe, S., et al.: Attention is not not explanation. In: EMNLP-IJCNLP (2019)
45. Wiegreffe, S., et al.: Measuring association between labels and free-text rationales. arXiv preprint arXiv:2010.12762 (2020)
46. Xue, L., et al.: mT5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934 (2020)