The impact of explanations in Natural Language Inference: an experimental analysis on T5 models

Andrea Roncoli, Bianca Ziliotto

February 2024

Github: https://github.com/DreRnc/ExplainingExplanations.

Abstract

This work explores the impact of explanations in fine-tuning a T5 model on a Natural Language Inference task. After measuring the model's ability to perform NLI, we investigate the use that T5 makes of explanations, and its capability to generate them. A set of experiments is designed to evaluate quantitatively and qualitatively the impact of generating explanations to justify the labels predicted.

1 Introduction

Explaining neural network predictions is a notoriously difficult problem, which has led to a body of research that aims at making neural networks more easily interpretable.

One of the proposed solutions for model interpretation in Natural Language Processing (NLP) sees the use of datasets annotated with explanations. Such explanations can be shown to the model as examples in few shots learning [3], or they can be used to fine-tune a model, training the model itself to output an explanation for each produced prediction [4].

In any case, providing the model with explanations has not only the benefit of making it more interpretable, but can also possibly enhance the model performance in terms of accuracy of the predictions, proving the model's capability of making a good use of explanations. The increase of accuracy due to explanations has already been shown in few-shot learning, where explanations are fed in input as examples. A different thing is understanding whether it is sufficient to add explanations to the outputs of training samples, forcing the model to generate explanations for its predictions, to obtain a change in the model's performance, which is the setting adopted in our experiments.

Here, we mainly focus on assessing the performance improvement, if any, allowed by explanations on a Natural Language Inference (NLI) task; at the same time, we monitor the ability of the model to generate plausible explanations for the choice of the label. Finally, we try to take a further step in understanding the role of explanations by giving the model compressed versions of the same annotations, with the aim of identifying the most significant features - semantic or syntactical - brought by explanations.

1.1 Model

In [5], a framework is defined to cast all NLP tasks as “text-to-text” problems, leading to the introduction of the Text-To-Text Transfer Transformer (T5) model. Such framework makes it particularly straightforward to train a model to produce explanations, since it is sufficient to concatenate explanations to the output texts. For this reason, all our experiments are executed on T5 models. In particular, in order to investigate the scalability of our results, we repeat all experiments on two different sizes of T5: T5-Small (60 million parameters) and T5-Base (220 million parameters).

1.2 Dataset

The dataset that we use for experiments is the Explained Stanford Natural Language Inference (E-SNLI) dataset [2], an English dataset for NLI extended with explanations. Notice that this dataset is independent from MultiNLI [7], on which T5 is already pre-trained; thus, the model is pre-trained on NLI task but on a different dataset (and with no explanations).

The dataset consists of a training set (549367 samples), a validation set (9842 samples) and a test set (9824 samples).

An example taken from E-SNLI dataset is shown in Table 1. As we can see, each sample contains a premise and a hypothesis: the label expresses the relationship (entailment, neutral or contradiction) between the premise and the hypothesis. The respective explanation provides further elements to understand and justify the choice of the label.

Premise	A person on a horse jumps over a broken down airplane.
Hypothesis	A person is training his horse for a competition.
Label	Neutral
Explanation	The person is not necessarily training his horse.

Table 1: Sample from E-SNLI dataset.

2 Methods

We perform a series of experiments with T5-Small and T5-Base in order to evaluate the increase of accuracy during fine tuning, with and without explanation generation, also looking at the scalability of the performance with the size of the model.

After assessing the performance of the pre-trained models (Section 2.1), we evaluate the improvement obtained by fine-tuning the models without using the explanations (Section 2.2); then, we teach the model to generate explanations, observing the differences in terms of accuracy of label prediction (Section 2.3).

Subsequently, we shuffle the explanations of different samples to measure the impact of a wrong annotation on the model's accuracy (Section 2.4). Finally, we extract the linguistic profile of annotations and we feed the model with a compressed version of explanations (Section 2.5).

2.1 Zero-shot evaluation of pre-trained models

We first test our models on the E-SNLI test set without any fine-tuning, measuring its zero-shot performance in terms of accuracy of the predicted labels. Since T5 is already trained on other NLI datasets, it is important to evaluate its performance and keep it as a reference for further improvement.

The model is given prompts in the form:

```
premise:<premise> hypothesis:<hypothesis>
```

and is given labels in text form: `[entailment, neutral, contradiction]`.

2.2 Fine-tuning without explanations

The second step consists in fine-tuning the models after removing explanations from the dataset, in order to set a new reference to compare with the following results to extract and quantify the contribution brought by explanations. As in all the following experiments, the validation set is used to understand when the model is overfitting the training set; the best model is selected as the model with the highest validation accuracy.

2.3 Fine-tuning with explanations

In this experiment, we add explanations to the labels, meaning that we teach the model to generate also the explanations. The labels are thus in the form:

```
label:<label> explanation:<explanation>
```

We also try to use labels in the inverted format (only in the T5-Small model, due to time and computational constraints):

```
explanation:<explanation> label:<label>
```

The accuracy, however, is measured only in terms of the predicted label (entailment, neutral, contradiction), ignoring the explanations. We seek to evaluate if forcing the output model to generate explanations for its answers improves the model's performance in predicting the label.

Furthermore, as in all the subsequent experiments with explanations, we monitor the correctness of the produced explanations: after encoding both true and predicted explanations using a Sentence BERT (SBERT) [6] model (all-MiniLM-L6-v2), we compute the cosine similarity among the encodings. The average cosine similarity provides a measure of the model's ability to predict explanations which are similar to the ground truth, and thus hopefully useful.

2.4 Fine-tuning with shuffled explanations

After assessing the effect of explanations, we are interested in measuring the impact of completely useless and possibly misleading explanations on the model's performance. This serves as a standard control experiment to asses if the model is actually using explanations in a semantically meaningful way. To investigate this aspect, we shuffle the training explanations of different samples, thus feeding the model with explanations that have no correlation with the premises and hypothesis.

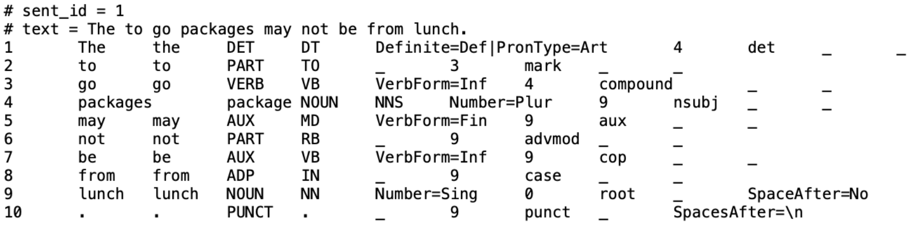
2.5 Fine-tuning with modified explanations

In case the model is not able to make a good use of explanations to boost the accuracy, we try to compress the information contained in explanations, seeing if this can make the task easier.

2.5.1 Text profiling and syntactic extraction

To help us in this task, we use Profiling-UD[1], a web-based application devised to carry out linguistic profiling of texts. In particular, we extract morphosyntactic information from the dataset explanations, and we use such information to manipulate the format of explanations in multiple ways.

Profiling-UD provides us with a set of features for each token: among these features, we can find the lemma (i.e. the base form of the token) and the universal part-of-speech (UPOS) tag, as can be seen in Figure 1.



```
# sent_id = 1
# text = The to go packages may not be from lunch.
1       The     the     DET     DT      Definite=Def|PronType=Art       4       det     _       _
2       to      to      PART    TO      _       3       mark    _       _
3       go      go      VERB    VB      VerbForm=Inf    4       compound        _       _
4       packages        package NOUN    NNS     Number=Plur     9       nsubj   _       _
5       may     may     AUX     MD      VerbForm=Fin    9       aux     _       _
6       not     not     PART    RB      _       9       advmod  _       _
7       be      be      AUX     VB      VerbForm=Inf    9       cop     _       _
8       from    from    ADP     IN      _       9       case    _       _
9       lunch   lunch   NOUN    NN      Number=Sing     0       root    _       SpaceAfter=No
10      .       .       PUNCT   .       _       9       punct   _       SpacesAfter=\n
```

Figure 1: Conllu standard representation of an explanation of the dataset, extracted with the Profiling-UD tool.

The experiment that we design consists in replacing the explanations with restricted lists of tokens, corresponding to nouns and verbs contained in the original explanations. In this way, we aim at understanding whether these tokens are sufficient for a good use of the explanations, while compressing the length and perhaps helping the model to take explanations into account while generating the labels. Due to time and computational constraints, this experiment is ran only on the T5-Small.

3 Results and discussion

The accuracies of the best models, selected through validation, on the test set are reported in Table 2.

	T5-Small	**T5-Base**
Pre-trained model	65.50%	80.34%
FT (labels)	**88.39%**	**91.01%**
FT (labels; explanations)	87.00%	90.75%
FT (explanations; labels)	83.19%	-
FT (labels; wrong explanations)	87.63%	90.84%
FT (labels; compressed explanations)	87.59%	-

Table 2: Accuracies obtained on the test set with the pretrained and fine-tuned (FT) models, with different experimental set-ups.

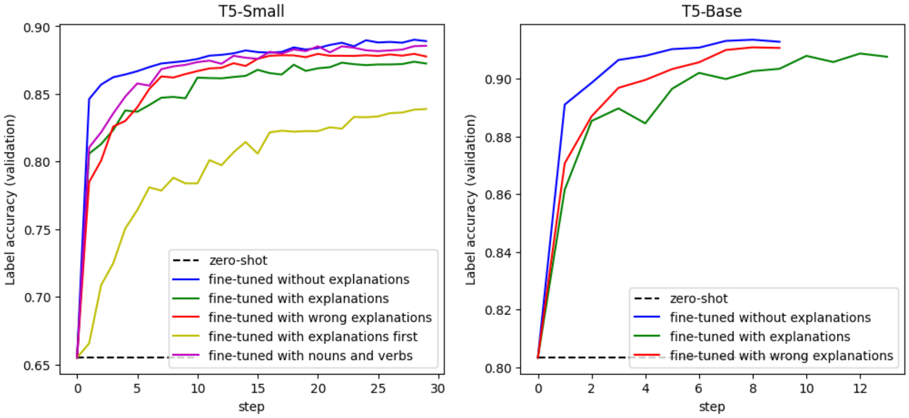


Figure 2: Accuracy in label prediction in the different experimental setups, for T5-Small and T5-Base. The evaluation steps (on the horizontal axis), are made every 5000 training steps (thus ≈ 3 times per epoch). T5-Base is trained for less steps as it tends to overfit faster (and it is also more expensive to train).

The first experiment (Section 2.2) shows that the accuracy of the pre-trained model can be further increased with few epochs of training: this confirms that the NLI dataset for which T5 was trained is independent from the E-SNLI corpus. In particular, 10 epochs of fine-tuning improve the accuracy of T5-Small from 65% top 88%; as for T5-Base, 3 epochs are sufficient to reach 91% (from 80% of the pre-trained model), and further fine-tuning results in over-fitting.

The second experiment (Section 2.3) does not seem to produce any improvement in the model's accuracy. What we notice in both T5-Small and T5-Base

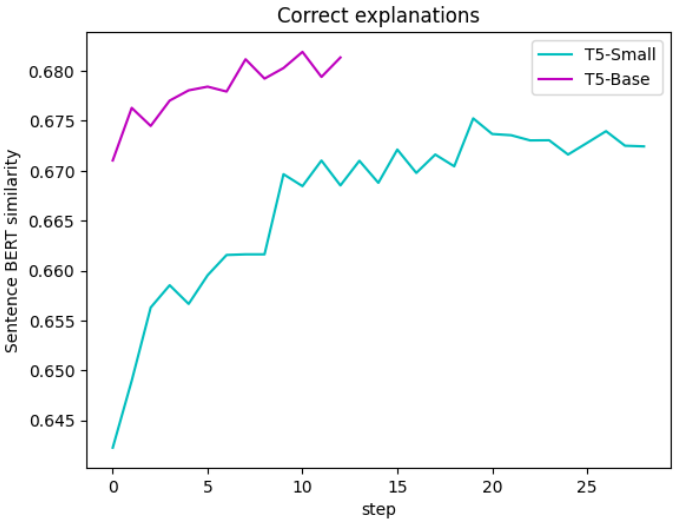


Figure 3: Cosine similarity between label and predicted explanations, encooded with SBERT.

is that the learning process requires more steps to reach the same accuracy, as seen in Figure 2, which suggests that the model is trying to improve the accuracy of label predictions and at the same time trying to generate better explanations. This intuition is confirmed by the increasing trend of the average similarity of SBERT (Figure 3). However, despite having shown that the model is learning to produce better explanations, we have no reason to conclude that learning explanations is contributing to a higher performance in the NLI task: on the contrary, the effort required to learn the two tasks together is slowing down the accuracy curve (Figure 2). Moreover, we observe that postponing the label to the explanation further slows down the learning process. Observing the outputs in the initial phase of the training, we suggest that this happens as the model takes more time to learn the correct format to output the explanations and labels, as labels are appended to sequences of variable length (the explanations).

The results of the third experiment (Section 2.4) allow us to consider the effect of wrong explanations. In this setting, the model achieves an extremely low (≈ 0.1) average cosine similarity between SBERT embeddings of generated and ground-truth explanations. This is to be expected, considering that it has been trained with random explanations. Furthermore, we notice that the accuracy of the model is not negatively affected by the wrong annotations (Figure 2). On the contrary, the learning process is actually sped up with respect to using the correct explanations. We conjecture that this is due to the fact that the model is simply learning to ignore the explanations, thus avoiding spending time on learning to generate correct explanations.

In the fourth experiment (Section 2.5), we observe that the learning process is even faster than in the third experiment, although slower than the case of no explanations being used. This may be explained by the fact that the model is again learning to simply ignore the explanations as it is not able to extract valuable information from such a compressed representation, and can do so faster simply because explanations are much shorter in this setting.

3.1 Qualitative evaluation

Observing the quantitative results obtained in Section **??**, we cannot conclude that this use of explanations can benefit the T5 accuracy in the NLI task. However, we are interested in deepening our understanding of what the T5 model is learning through a qualitative analysis of generated explanations. First of all, we want to see whether the generated explanations are syntactically correct and semantically meaningful. Also, since the model obtains a very high accuracy on NLI, we are interested in studying the cases where it gives wrong answers, seeing if we can get an intuitive motivation for the mistakes made by T5.

Premise	**Hypothesis**	**True Label**	**T5 Label**	**T5 Explanation**
A land rover is being driven across a river.	A vehicle is crossing a river.	E	E	A land rover is a vehicle.
Several women are playing volleyball.	The women are hitting a ball with their arms.	E	N	Just because women are playing volleyball does not mean they are hitting a ball with their arms.
This church choir sings to the masses as they sing joyous songs from the book at a church.	The church has cracks in the ceiling.	N	C	The church choir cannot sing joyous songs from the book at a church if it has cracks in the ceiling.
Three boys in white shirts are walking behind an older model Ford car.	The boys are on the street	E	N	Just because the boys are walking behind a car does not mean they are on the street.

Table 3: Examples of predicted labels (E: Entailment; N: Neutral; C: Contradiction) with the respective generated explanations.

The first example in Table 3 shows the ability of T5 to justify its correct predictions on the base of known external information, i.e. 'a land rover is a vehicle'. In the following example, instead, T5 seems to lack this same ability, failing to understand the semantic implications of playing volleyball. The third case is slightly different and very interesting. The model sees a contradiction between a positive (joyous songs) and a negative event (cracks in the ceiling), but shows no common sense in understanding the proportions between these things. Cracks in the ceiling, in fact, would never prevent the joyous choirs despite being an inconvenient: this is obvious to a human, but T5 seems to lack this kind of common sense. Finally, we show a case where the mistake of T5 is due to an ambiguity in the true label. The fact that children are walking behind a car could make us think that they are on a street, but this is not strictly necessary, thus we think that many humans would disagree on the true label.

Interestingly, we notice that in all shown cases (and also in the other $\approx$ 100 cases that we analyzed), the logical structure of the explanation is always coherent with the choice of the label, may it be right or wrong. For example, the explanation for a neutral label is many times expressed in the form "just because ¡premise¿ it doesn't mean that ¡hypothesis¿". This seems to suggest that whilst T5 sometimes struggles with semantics (e.g. integrating external information), it learns the syntactical and logical structures very well.

After evaluating the quality of explanations generated by T5 trained on the correct explanations, we are also interested to analyze the models fine-tuned with wrong or compressed explanations. In both cases, we can validate our intuition about explanations being ignored: in fact, T5 tends to produce the same explanation for each test sample, without making any effort to match the context of the premise and hypothesis. For example, the model trained with compressed explanations curiously generates the words "man woman genders" at the beginning of every explanation (eventually followed by other senseless words).

References

[1] Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. Profiling-UD: a tool for linguistic profiling of texts. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7145–7151, Marseille, France, May 2020. European Language Resources Association.

[2] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations, 2018.

[3] Andrew K. Lampinen, Ishita Dasgupta, Stephanie C. Y. Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L. McClelland, Jane X. Wang, and Felix Hill. Can language models learn from explanations in context?, 2022.

[4] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. Wt5?! training text-to-text models to explain their predictions, 2020.

[5] 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, 2023.

[6] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[7] Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference, 2018.