Robot Explains Robot how to Explain Natural Language Inference

Thijmen Dam **Folkert Leistra** **Ludwig Sickert**

t.m.dam@student.rug.nl, f.a.leistra@student.rug.nl, l.m.sickert@student.rug.nl

University of Groningen
Groningen, The Netherlands
GitHub: https://github.com/lsickert/ik-nlp2022-group11

Abstract

Over the past years there has been an increasing effort to make the decision processes of neural networks more transparent. In the field of Natural Language Processing this is often done by adding machine-generated explanations to the output of a given language model. Camburu et al. (2018) explore different ways to apply this principle to the task of NLI, which we are expanding upon in this paper. We find that there are several approaches that are able to generate explanations while performing similar or better than traditional NLI models, namely the joint prediction of labels and explanations and the prediction of labels from the generated explanations, implicating that it might be favorable to use these approaches instead of pure classification models in the future.

1 Introduction

The way Neural Networks derive their predictions is generally considered a "black-box" by researchers in the field of Artificial Intelligence. In recent years there has therefore been an increased emphasis on designing models that are able to explain precisely how they arrived at their prediction in order to increase the transparency of the decision process.

In the field of Natural Language Processing (NLP) in particular, a growing number of works propose to solve this problem by adding modules to a model that can generate natural language explanations about their predictions (Camburu et al., 2020; Liu et al., 2019a; Park et al., 2018).

One particular task where explanations are interesting is that of Natural Language Inference (NLI), which tries to predict whether a premise *entails*, *contradicts* or is *neutral* towards a given hypothesis.

In this paper, we explore how NLI tasks can be expanded with machine-generated explanations to provide a deeper insight into the decision process of the neural network. To do this, we expand on the work by Camburu et al. (2018), which started exploring the importance of explanations in these kinds of tasks with a series of experiments.

1.1 The Team

Our team consists of Folkert, Ludwig and Thijmen. We have worked closely together in the beginning stage by drafting a clear approach for this project. After deciding on our approach, we divided the project into sub-tasks. Finally, we worked together to create our final report and codebase.

Folkert is an Information Science (MSc) student with a background mainly focused on NLP techniques. During this project, Folkert has mostly worked on the Predict and Explain task and the coordination of the project. As working with an encoder-decoder model was a new experience, Folkert was keen to start working on the generation of natural language explanations.

Ludwig is an Artificial Intelligence (MSc) student with a practical background in NLP and using Artificial Intelligence on unstructured data. During the project, he focused on training the classification models and the Explain from Classification experiment. He also optimized the scripts to be run on the Peregrine high performance computing cluster.

Thijmen is an Information Science (MSc) student with a background in NLP and development. Thijmen has been a flexible participant aiding others when needed during this project. Thijmen worked on the Predict and Explain model, performed a data analysis on the results, and set up the error analysis.

2 Experimental Setup

In this section, we will elaborate on the experiments that we have conducted and the data that was used to complete these experiments.

2.1 Data

We used the e-SNLI dataset created by Camburu et al. (2018), which consists of roughly 570,000 examples containing a premise and hypothesis together with a label corresponding to the three classes *entailment*, *neutral* and *contradiction*. The data is already split into separate train, test and validation sets, with the test and validation set containing 9842 samples each and the training set containing the rest of the samples. In addition to the parameters mentioned above, the three sets also contain human-created explanations for the labels, with the training set containing one explanation per example and the test and validation set containing three explanations per sample.

Before working with our models, we had to identify the maximum input - and output length for our models. For our maximum input length, we concatenated the premise and hypothesis, separated by a [SEP] token, and used the BART-base tokenizer (Lewis et al., 2019) to obtain the maximum length of the input-ids. We performed a similar process to determine the maximum output length but now for the concatenation of the label and the first explanation, using a space as a separator. This method resulted in a maximum input length of 125 tokens and a maximum output length of 193 tokens.

To compare the quality of the explanations generated by our model, we use the ROUGE (Lin, 2004), and BARTScore (Yuan et al., 2021) scores. In addition, to measure the effectiveness of the labels predicted by our model, we will calculate the Accuracy as well as Precision, Recall and F1-Scores.

2.2 Baseline

To first obtain a baseline on the results, we fine-tuned a RoBERTa model in the standard configuration described in Liu et al. (2019b) to classify the dataset. As input, we used a Byte-Pair encoded (Sennrich et al., 2016) concatenation of the premise and hypothesis with a learning rate of 5e-5, Adam with Weight Decay (Loshchilov and Hutter, 2017) as an optimization function and a linear learning rate decay. We trained the model for three epochs with a batch size of 32.

2.3 Explain from Classification

Since we ideally want to generate explanations directly from the classifying model without altering its internal states, we used the last hidden state from the baseline model as input to fine-tune another RoBERTa model with a Causal Language Modeling head with the tokenized human-created explanation as the label. The model was again trained with the hyperparameters above, but this time with two epochs and a batch size of 64.

2.4 Predict and Explain

Our second experiment consisted of using two encoder-decoder models to jointly predict the NLI label and the explanation for that label. The input for these models was the same as for the baseline. The models were trained on the tokenized concatenation of the NLI label and the first explanation. The label and explanation were concatenated by adding a space. The first model that we used was the BART-base model (Lewis et al., 2019). For our second model, we used the Text-to-Text Transfer Transformer (T5-base) model (Raffel et al., 2019).

Both models were trained using the seq2seq trainer from HuggingFace. We did not perform an extensive hyperparameter optimization search and instead used slightly adapted default settings. For both models, we used a batch size of 16 for 1 epoch, an Adam optimizer with a learning rate of 2e-05 and a weight decay of 0.01.

2.5 Predict using Explanations

As a final experiment, we fine-tuned two RoBERTa models with the same hyperparameters as the baseline on the human-generated explanations as well as the machine-generated explanations from the previously fine-tuned BART-model to see if the NLI task could be solved with an input of only explanations. This is an approach that was also taken by Camburu et al. (2018), which led to good results in their paper.

2.6 Technical Setup

To train and evaluate our models, we used the PyTorch (Paszke et al., 2019) and Transformers

(Wolf et al., 2020) libraries, from where we also obtained the pre-trained baselines for our finetuned models. To obtain the ROUGE (Lin, 2004) and BARTScore (Yuan et al., 2021) metrics, we used the respective libraries published by the authors.

The code used for the experiments in this paper as well as a link to the trained models can be found under `https://github.com/lsickert/ik-nlp2022-group11`.

3 Results

3.1 Baseline

The baseline RoBERTa model achieved an accuracy of 85%, which is in line with the performance of the authors of the e-SNLI dataset (Camburu et al., 2018). The remaining performance metrics are shown in Table 1 below.

	Prec.	**Rec.**	**F1**	**Supp.**
entailment	0.78	**0.93**	0.84	3368
neutral	0.85	0.75	0.80	3219
contradiction	**0.94**	0.86	**0.90**	3237
accuracy			0.85	9824
macro avg	0.86	0.85	0.85	9824
weighted avg	0.86	0.85	0.85	9824

Table 1: Classification metrics for the baseline RoBERTa model.

3.2 Explain from Classification

Unfortunately, the approach of generating explanations directly from the hidden states of the baseline classification model did not lead to good results, as shown in the examples (1) and (2). We refrain from reporting the ROUGE and BARTScore metrics for this approach since it can intuitively be seen that the generated sentences are not very useful.

(1) Entailment: The man is is a a

(2) Contradiction: Just because is not a a a a

We cross-validated these results with several training runs with different configurations for learning rate(5e-4 - 1e-7, weight decay (0 - 1e-5) and epochs (2-10), so we can conclude that they are somewhat stable.

After performing additional qualitative analysis on the results, we could, however, see that the model has at least learned to generate slightly different explanations for *neutral*, *entailing* or *contradicting* classifications as showcased by the examples above.

3.3 Predict and Explain

For the Predict and Explain sub-task, we have decided to include several evaluation metrics. For the evaluation of the NLI label prediction, we have generated classification reports, which can be found in Table 2 and Table 3. The scores that we obtained for both the BART and T5 models are reasonably impressive. The BART model slightly outperforms the T5 model by obtaining an F1 score of 0.90 and an accuracy of 0.88. However, the scores of the T5 model are very similar, with an F1 score of 0.89 and an accuracy of 0.87.

For the evaluation of the generated explanations, we have used one text-based metric by using the ROUGE score and one neural metric by using the BartScore. Table 4 shows the BartScores that our models have obtained. With the BartScore, we are able to quantify the similarity between the model generated explanations and the gold explanations. To analyze the BartScores, we have decided to include the mean, median, minimum and maximum BartScore. Furthermore, we analyze nine sentences per model based on the BartScore in section 3.3. From the BartScores, we can conclude that the BART model slightly outperforms the T5 model in generating the explanations. However, the BartScores are very similar, and the T5 model does obtain a smaller minimum BartScore. Additionally, we have included the distribution of the BartScores for both models, which can be found in Figure 1 and Figure 2. These graphs show us that the distribution is reasonably in line with the median BartScores.

The final metric that we used to evaluate generated explanations are the ROUGE scores, which can be found in Table 5. The table shows us that the BART model is able to outperform the T5 model slightly on all ROUGE variants. However, it is interesting to see that also, for this metric, the scores are very similar.

3.4 Error Analysis

To better understand the behaviour of our Predict and Explain models, we additionally performed an qualitative error analysis. Per model, we ana-

	Prec.	**Rec.**	**F1**	**Supp.**
entailment	0.91	0.88	**0.90**	3368
neutral	0.83	0.88	0.85	3219
contradiction	**0.92**	**0.89**	**0.90**	3237
accuracy			0.88	9824
macro avg	0.88	0.88	0.88	9824
weighted avg	0.88	0.88	0.88	9824

Table 2: BART classification report for the Predict and Explain task.

	Prec.	**Rec.**	**F1**	**Supp.**
entailment	**0.90**	0.88	**0.89**	3368
neutral	0.83	0.83	0.83	3219
contradiction	0.87	**0.89**	0.88	3237
accuracy			0.87	9824
macro avg	0.87	0.87	0.87	9824
weighted avg	0.87	0.87	0.87	9824

Table 3: T5 classification report for the Predict and Explain task.

	Mean	**Median**	**Min**	**Max**
BART	**-2.023**	**-1.967**	-5.872	**-0.131**
T5	-2.059	-2.023	**-5.654**	-0.172

Table 4: BART scores for BART and T5.

	R1	**R2**	**R-L**	**R-L SUM**
BART	**48.99**	**23.94**	**43.74**	**43.92**
T5	48.60	23.44	43.46	43.62

Table 5: ROUGE scores for BART and T5. The table displays the ROUGE1 (R1), ROUGE2 (R2), ROUGE-L (R-L) and ROUGE-L SUM (R-L SUM).

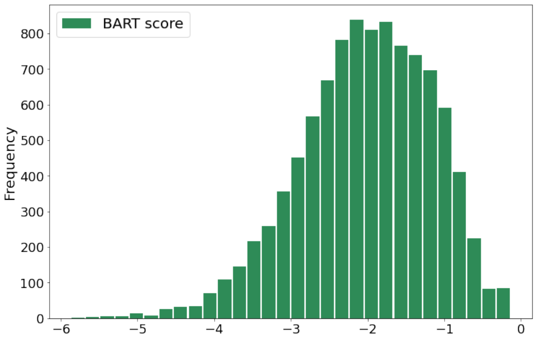

Figure 1: BART scores distribution for the BART model.

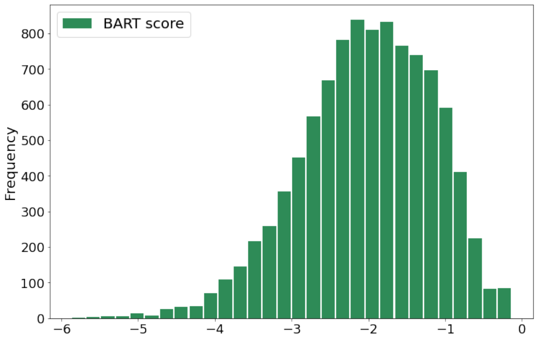

Figure 2: BART scores distribution for the T5 model.

lyzed nine generations which can be found in the appendix of this report (Table 8 and Table 9). We have chosen the generations to analyze based on the BartScore: We choose three generations that score the highest, three generations around the median value and three generations that score the lowest. By doing so, we are able to quickly analyze some of the errors the models make. The tables include the premises, hypothesis, gold explanation, model explanation and our analysis. In the section below, we will list some of the most common errors that we have found:

- The generated explanation is incomplete or cut-off early.
- The generated explanation is too long.
- The generated explanation does not follow correct grammatical sentence structure.
- The generated explanation uses the wrong aspects of the premise and hypothesis.
- The generated label is incorrect, and the explanation follows that template.
- The generated explanation is correct but differs drastically from the gold explanation.

While the first three errors can easily be solved by further optimization of the model parameters, the next two might be harder solve, since they touch upon some of the general issues that current language models often have like the notion of the "stochastic parrot" (Bender et al., 2021) and the associated lack of real understanding these models have.

Finally, we would like to add that both the BartScore and ROUGE scores are not entirely optimal for use as an evaluation metric. While they give a pretty good indication of the correctness and similarity of the generated natural language, the generated explanation can still be correct even though the BartScore is very low. This has to do with different wording choices or ambiguity in the relation between the premises and hypothesis. The last observed error is therefore less an error with the language model, but inherent to the evaluation mechanism.

3.5 Predict using Explanations

The model using human-generated explanations achieved a very high accuracy of 98%, which implies that an NLI classification without using the actual premise and hypothesis, but only the explanations, is not only feasible but outperforms the classical model shown in Section 3.1. The complete overview of the performance metrics is shown in Table 6.

	Prec.	**Rec.**	**F1**	**Supp.**
entailment	0.98	0.98	0.98	3368
neutral	0.98	0.97	0.97	3219
contradiction	0.98	0.98	0.98	3237
accuracy			0.98	9824
macro avg	0.98	0.98	0.98	9824
weighted avg	0.98	0.98	0.98	9824

Table 6: Classification metrics for the RoBERTa model trained on human-generated explanations.

Since the model performed well using the human-generated explanations as input, a natural extension was it then to use the machine-generated explanations as well. This model achieved a still high accuracy of 83%, which is worse than the human-generated explanations but still almost as good as our baseline model. The full performance metrics are shown in Table 7.

	Prec.	**Rec.**	**F1**	**Supp.**
entailment	0.81	0.98	0.89	3368
neutral	0.95	0.66	0.78	3219
contradiction	0.79	0.86	0.82	3237
accuracy			0.83	9824
macro avg	0.85	0.83	0.83	9824
weighted avg	0.85	0.83	0.83	9824

Table 7: Classification metrics for the RoBERTa model trained on machine-generated explanations.

4 Conclusions

This paper explored the feasibility of generating explanations to make the decision process in NLI tasks more transparent. We expanded on the work by Camburu et al. (2018) with a series of experiments.

The results clearly show that using explanations as input for NLI tasks is feasible and can outperform classical models based on the premise and hypothesis. However, since it is not always feasible to obtain human-generated explanations for the task at hand, we showed that auto-generating explanations from the premise and hypothesis through a sequence-to-sequence model might be a good alternative. Under these models, we saw that the joint prediction of class labels and generation of explanations through the same model performed the best, with an accuracy of 90%. However, the decoupled explanation and classification models also performed decently, and it is very likely that this approach could also be brought to similar performance levels with additional optimization. It, therefore, can also be seen as an effective way to perform NLI, for example in situations where a partially labelled dataset is available or the decoupling of explanation generation and NLI classification is necessary.

The poor performance of the generative model using the classification hidden states as input was unexpected but seemed to be relatively stable across different hyperparameter configurations. We still believe that this approach would ultimately be the most preferable since it is done without altering the original NLI classification model or task and could therefore also be applied to existing models. Further research might be warranted to confirm the (in)feasibility of this approach.

Finally we would like to add that the usefulness of these machine-generated explanations is still something that needs to be better understood and explored since, as shown by our error analysis, it could also be argued that the model is not actually describing its inner workings but merely optimizing for an additional task, the generation of a matching explanation to a label.

Acknowledgements

We thank the Center for Information Technology of the University of Groningen for their support and

for providing access to the Peregrine high performance computing cluster.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Hui Liu, Qingyu Yin, and William Yang Wang. 2019a. Towards explainable NLP: A generative explanation framework for text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5570–5581, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *CoRR*, abs/2106.11520.

A Model Output Analysis (BART)

BEST 3	
PREMISE	An Indian woman is dancing with her partner.
HYPOTHESIS	A woman is moving.
GOLD	0 Dancing is a form of moving.
GENERATED	0 Dancing is a form of moving.
BART SCORE	-0.131
ANALYSIS	The gold explanation and generated explanation are identical.
PREMISE	Three women sit at a table, taking notes from three identical magazines.
HYPOTHESIS	Three women are reading and writing.
GOLD	0 Taking notes implies reading and writing.
GENERATED	0 Taking notes implies reading and writing.
BART SCORE	-0.144
ANALYSIS	The gold explanation and generated explanation are identical.
PREMISE	Three puppies are in the tub being sprayed with water by a person.
HYPOTHESIS	Animals are in a tub.
GOLD	0 Puppies are animals.
GENERATED	0 Puppies are animals.
BART SCORE	-0.151
ANALYSIS	The gold explanation and generated explanation are identical.
MEDIAN	
PREMISE	A boy in swimming trunks wades out into the water.
HYPOTHESIS	The young swimmer wades out into the water at the start of his team practice.
GOLD	1 Just because a boy is swimming does not mean he is at his team practice.
GENERATED	1 A boy in swimming trunks wades out into the water does not indicate that
BART SCORE	-1.967
ANALYSIS	The generated explanation is quite different to the gold explanation, but not incorrect. Unfortunately, the explanation is cut short, otherwise it could have been correct.
PREMISE	A man holding some sort of vegetables while his wife in a red sweater stares to the side, and his daughter in a red sweater looks at the camera.
HYPOTHESIS	A family poses for a Christmas picture.
GOLD	1 Just because they’re posing for a picture doesn’t mean they’re posing for a Christmas picture.
GENERATED	1 A man holding some sort of vegetables while his wife in a red sweater stares to
BART SCORE	-1.967
ANALYSIS	This generated explanation is completely wrong. However, the model did predict the label correctly. In this case, the model just copies the premise as explanation.
PREMISE	Girl wearing white shirt sings on stage while playing guitar
HYPOTHESIS	A girl plays guitar and sings for a crowd.
GOLD	1 A girl can sing or play guitar on a stage that doesn’t have a crowd.
GENERATED	1 Just because a girl sings on stage while playing guitar does not mean she sings for
BART SCORE	-1.967
ANALYSIS	The generated explanation is quite different to the gold explanation, but not incorrect. Unfortunately, the explanation is cut short, otherwise it could have been correct.
WORST 3	
PREMISE	A man smokes while leaning against a pillar, which has a ’missing person’ flier posted on it.
HYPOTHESIS	a guy smoking against a wall with a flier on it
GOLD	0 Pillar has ’missing person’ posted implies a flier on it.
GENERATED	0 A pillar is a wall.
BART SCORE	-5.675
ANALYSIS	The generated explanation is very different as the gold explanation. However, the model does classify correctly as an entailment and the explanation is in line with this choice. However, the generated explanation misses the key aspects.
PREMISE	A gray rally car is being driven by two adult males while a large white container is tied down on top of the roof.
HYPOTHESIS	The men hope the container won’t fly off the roof.
GOLD	1 How can a countainer fly off from a relly car
GENERATED	1 Sentence 1: A gray rally car is being driven by two adult males while
BART SCORE	-5.726
ANALYSIS	The label is correct, but the generated explanation is completely wrong.
PREMISE	A woman in a black leather shirt trench-coat and black leather calf-high boots and carrying a colorful bag with a shoulder strap faces away from the camera with a yellow post visible just in front of her.
HYPOTHESIS	Some stars are walking past the paparazzis and getting into an awards building
GOLD	1 beautifully dressed woman are not always stars, men with camera and yellow post are not paparazzis and the ceremony building always not awards building
GENERATED	2 A woman is not some stars.
BART SCORE	-5.872
ANALYSIS	The generated explanation is not the same as the gold explanation. However, the model does assign this case as a contradiction and the explanation is in line with this choice.

Table 8: Output analysis of our BART model. Examples are automatically selected based on their BART score.

B Model Output Analysis (T5)

BEST 3	
PREMISE	Two racers race on motorcycles.
HYPOTHESIS	Two people are racing.
GOLD	0 Racers are people.
GENERATED	0 Racers are people.
BART SCORE	-0.172
ANALYSIS	The gold explanation and generated explanation are identical.
PREMISE	Villagers pulling in the fish boat.
HYPOTHESIS	There are people pulling a boat.
GOLD	0 Villagers are people.
GENERATED	0 Villagers are people.
BART SCORE	-0.194
ANALYSIS	The gold explanation and generated explanation are identical.
PREMISE	A man and a boy behind the wheel of a car.
HYPOTHESIS	The man and the boy are buying shoes.
GOLD	2 The man and boy cannot be behind the wheel of a car and buying shoes at the same time.
GENERATED	2 The man and boy cannot be behind the wheel of a car and buying shoes at the
BART SCORE	-0.194
ANALYSIS	The gold explanation and generated explanation are identical.
MEDIAN	
PREMISE	Two adult men in black t-shirt and loud tunics hold smoking torches in preparation for an act of entertainment or celebration in a forested environment.
HYPOTHESIS	Two brothers hold torches to celebrate.
GOLD	1 Just because two men are preparing for a celebration doesn’t mean they are brothers
GENERATED	1 Not all men are brothers.
BART SCORE	-2.023
ANALYSIS	Even though the gold explanation and the generated are different, the generated explanation is not wrong. However, it is a little bit short.
PREMISE	Four children are wearing blue hats with colored streamers.
HYPOTHESIS	Four kids are waiting in line for the bus.
GOLD	1 children are wearing blue hats with colored streamers doesn’t always mean kids are waiting in line for the bus.
GENERATED	1 Just because the kids are wearing blue hats with colored streamers does not mean they
BART SCORE	-2.023
ANALYSIS	It looks like the generated explanation was going to be correct, but it is cut off early.
PREMISE	A group of children playing with props
HYPOTHESIS	The children are playing with machine guns.
GOLD	1 Props are not limited to machine guns.
GENERATED	2 The children cannot be playing with props and machine guns at the same time.
BART SCORE	-2.023
ANALYSIS	The generated explanation is not the same as the gold explanation. However, the model predicted this case to be a contradiction and did provide the right explanation for a contradiction case.
WORST 3	
PREMISE	The farmers are out working hard on there tractor, they have on their overalls and the machine is almost fixed.
HYPOTHESIS	Overall wearing people are working to fix a tractor.
GOLD	0 ’have on their overalls rephrase overall wearing and machine is fixed implies that they fix a tractor.
GENERATED	0 Farmers are people.
BART SCORE	-5.544
ANALYSIS	The generated explanation is different than the gold explanation. However, the model does predict the label correctly and the explanation is in line with entailment style explanations. Unfortunately, the aspects that are important in the explanation are missing.
PREMISE	A gray rally car is being driven by two adult males while a large white container is tied down on top of the roof.
HYPOTHESIS	The men hope the container won’t fly off the roof.
GOLD	1 How can a countainer fly off from a relly car
GENERATED	1 Just because a container is tied down on top of the roof does not mean it won
BART SCORE	-5.596
ANALYSIS	This generated explanation is not correct. The label is correct, but the explanation style belongs to contradictions. Moreover, the explanation is cut short.
PREMISE	A woman in a black leather shirt trench-coat and black leather calf-high boots and carrying a colorful bag with a shoulder strap faces away from the camera with a yellow post visible just in front of her.
HYPOTHESIS	Some stars are walking past the paparazzis and getting into an awards building
GOLD	1 beautifully dressed woman are not always stars, men with camera and yellow post are not paparazzis and the ceremony building always not awards building
GENERATED	2 A woman is not stars.
BART SCORE	-5.654
ANALYSIS	The generated explanation is not the same as the gold explanation. However, the model does assign this case as a contradiction and the explanation is in line with this choice.

Table 9: Output analysis of our T5 model. Examples are automatically selected based on their BART score.