Investigating the Effect of Natural Language Explanations on Out-of-Distribution Generalization in Few-shot NLI

Yangqiaoyu Zhou
University of Chicago
zhouy1@uchicago.edu

Chenhao Tan
University of Chicago
chenhao@uchicago.edu

Abstract

Although neural models have shown strong performance in datasets such as SNLI, they lack the ability to generalize out-of-distribution (OOD). In this work, we formulate a few-shot learning setup and examine the effects of natural language explanations on OOD generalization. We leverage the templates in the *HANS* dataset and construct templated natural language explanations for each template. Although generated explanations show competitive BLEU scores against groundtruth explanations, they fail to improve prediction performance. We further show that generated explanations often hallucinate information and miss key elements that indicate the label.

1 Introduction

Thanks to recent advances in pre-trained language models (Vaswani et al., 2017; Devlin et al., 2018), the state-of-the-art accuracy for natural language inference (NLI) can easily exceed 90% (Pilault et al., 2020). However, these NLI models show poor out-of-distribution (OOD) generalization. For instance, McCoy et al. (2019) create a templated dataset (*HANS*) and find model performance to be about chance in this dataset.

While recent studies try to tackle this robustness problem from the perspectives of both the dataset and the model (Le Bras et al., 2020; Swayamdipta et al., 2020; Clark et al., 2019), we investigate an extra dimension of information, natural language explanations. Our work is motivated by the growing interest in explanations in the NLP community (Camburu et al., 2018; Rajani et al., 2019; Alhindi et al., 2018; Stammbach and Ash, 2020): these explanations can potentially enable models to understand the reasoning strategy beyond spurious patterns. We focus on a few-shot learning setup because it is unrealistic to expect a large number of annotated OOD examples.

To introduce an OOD setting with natural language explanations, we construct E-HANS, a dataset with natural language explanations for each template in *HANS*. By leveraging the templates in *HANS*, we avoid the challenges in crowdsourcing natural language explanations (Wiegreffe and Marasović, 2021) and manually build an explanation dataset of high-quality.

We use an EXPLAINTHENPREDICT framework to learn with explanations. An explanation generation model outputs an explanation for each input example, and the generated explanation is fed into a classifier along with the input example. While BLEU scores imply high quality of generated explanations, learning with explanations does not improve predictive performance either in-distribution or out-of-distribution. We show the generated explanations contain words in the true explanations, but they fail to reproduce important phrases and often hallucinate entities during generation.

2 Building Natural Language Explanations for *HANS*

To investigate whether natural language explanations improve the robustness of natural language inference (NLI), we build on two existing datasets: 1) *HANS*, which introduces templates to generate OOD examples for robust evaluation of NLI models; 2) E-SNLI, which provides explanations for the Stanford Natural Language Inference (SNLI) dataset. Our key contribution is to augment *HANS* by building templated natural language explanations and studying the effect of these explanations on model robustness in a few-shot learning setup. Our dataset is available at https://github.com/ChicagoHAI/hans-explanations.

2.1 Existing Datasets

We start by presenting details of existing datasets.

***HANS* (McCoy et al., 2019)** contains NLI examples designed to be challenging for models that

Proceedings of the Second Workshop on Insights from Negative Results in NLP, pages 117–124

ISBN 978-1-954085-93-0

Premise: the psychologist by the programmers saw the essayist. **Hypothesis**: the psychologist saw the essayist. **Explanation**: the psychologist by the programmers is still the psychologist.

Table 1: An example from E-HANS. Average length of premise and hypothesis are 8.8 and 4.4 tokens. Average length of natural language explanation is 13.3 tokens.

tend to learn spurious patterns. It targets known heuristics for the majority of existing NLI data. For example, one heuristic assumes that a premise entails all hypotheses that are constructed using only words in the premise. There are 3 heuristics in *HANS*, each containing 10 subcases. A subcase is supported by a few templates and the dataset is constructed following these templates.

E-SNLI (Camburu et al., 2018) develops free-form self-contained explanations for the true labels in SNLI using crowdsourcing. We pretrain a model on this dataset to examine the effect of pretraining. There are three explanations collected for each example in the validation dataset, and we use the first explanation. We do not use the test set of E-SNLI.

2.2 Templated Natural Language Explanations for *HANS*

We build natural language explanations for *HANS* to examine whether explanations can help models when facing this challenging corpus. As *HANS* is constructed with templates, we develop templates for natural language explanations accordingly. They explain the reasons for the true label in human language.

Table 1 shows an example of the proposed explanations (more examples are in Appendix B). In addition to developing these templated explanations, we expand the original HANS vocabulary in terms of its nouns, verbs, adjectives, and adverbs to increase difficulty. This allows us to examine the effect of unseen words.

3 Experiments

3.1 Few-Shot Learning Set-up

To investigate whether natural language explanations improve the robustness of NLI models, we look at a few-shot learning setting. We focus on this setting since in practice one may have little or no access to OOD instances. We are interested in the following questions:

1. whether the model trained from in-distribution examples can generalize to unseen templates and words,
2. how many samples are enough for learning,
3. whether pretraining on E-SNLI improves generalization on *HANS*,
4. and most importantly, what is the effect of explanations.

We use 5-fold cross validation by splitting 118 templates randomly into 5 folds. We generate k samples for each training template using E-HANS explanation templates. We then build a corresponding development set that contains $0.2k$ samples of each training template, so that the development set is 20% the size of the training set and does not include any unseen template. This setup ensures that the size of the development set is realistic (Kann et al., 2019).

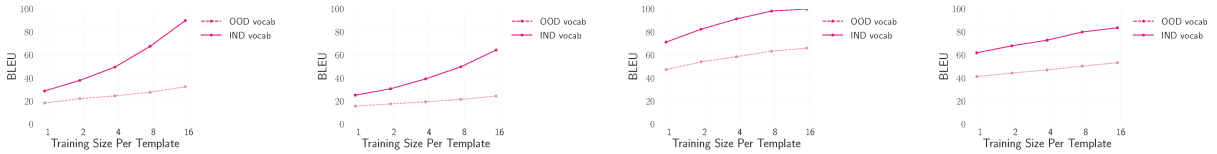
Finally, we build test instances in the following categories to evaluate the performance of the models both in-distribution and out-of-distribution:

- *IND vocab, IND template*. Both the templates and the vocabulary are matched with the training set. We expect the performance to grow steadily as k increases.
- *OOD vocab, IND template*. We use the same templates as the training set, but use unseen words to generate this test set. The challenge lies in understanding unseen words.
- *IND vocab, OOD template*. We use the unseen templates and the same vocabulary as the training set. The challenge lies in understand the logic encoded in unseen templates.
- *OOD vocab, OOD template*. Finally, we generate the test data with both the unseen templates and the unseen words.

We use the same test sets (300 examples for each template) to examine how the models' performance changes as k increases.

3.2 Models

We adapt the EXPLAINTHENPREDICT architecture introduced by Camburu et al. (2018) in our experiments. It consists of **a generation model** and **a classification model**. The generation model produces an explanation given an input premise and hypothesis pair. This generated explanation and the original input are fed into the classifier for label prediction. Our framework slightly differs from Camburu et al. (2018) in that their classifier only takes explanation as input for the classifier.



(a) IND templates with BERT (b) OOD templates with BERT (c) IND templates with ESNLI (d) OOD templates with ESNLI

Figure 1: $x-$axis shows the number of samples per template, while $y-$axis shows the BLEU score. BLEU scores are high for *IND vocab, IND template* instances. Although BLEU drops substantially for both BERT and the E-SNLI pretrained model under OOD vocab and OOD templates, it is still decent (above 40 with E-SNLI).

The explanation generation model follows an encoder-decoder framework. Both the encoder and the decoder use the BERT model, but the decoder uses a masking mechanism so that it predicts the next word considering only all the preceding words in both training and testing phases. Our generation model obtains close to SoTA performance on e-SNLI, comparing against WT5 (33.15 vs. 33.7 in BLEU) (Narang et al., 2020).

The classification model is a BERT sequence classifier, where a linear layer is applied to the pooled output of BERT encodings (i.e., embedding of the CLS token).

3.3 Experimental Setup

We used $k = 1, 2, 4, 8, 16$ to generate the training data. The explanation generator trains on groundtruth explanations. For all of our models, we saved the model with the best validation performance during training and did not tune other hyperparameters. We used a training batch size 16 and a learning rate 5e-5 for the explanation generator, and we used a training batch size 128 and learning rate 2e-5 for the classifier.

Model comparisons. To test if explanations help with learning, we compare with a baseline that only includes the classifier component with the input premise, hypothesis pair (hence "*label-only*"). We also consider a baseline that does not update with the k samples in the training set (hence "*no training*") and a majority baseline ("*majority*").

In addition, we compare the vanilla BERT model with a BERT model fine-tuned on E-SNLI during both generation and classification to investigate whether pretraining on E-SNLI helps with the *HANS* task.

4 Results

We first look at the quality of generated explanations using BLEU. Although the generated explanations match groundtruth explanations well based on BLEU, they affect downstream classification negatively in our few-shot learning set-up. We further examine the generated explanations to understand why the predictive performance drops when adding natural language explanations.

4.1 Quality of Explanations based on BLEU

Generated explanations achieve high BLEU scores based on our groundtruth templated explanations (Figure 1). IND vocab explanations can achieve BLEU scores greater than 90 on IND templates and 60 on OOD templates when $k = 16$. Even OOD vocab explanations can achieve BLEU scores close to 20. In general, the performance grows steadily as k increases for both IND and OOD.

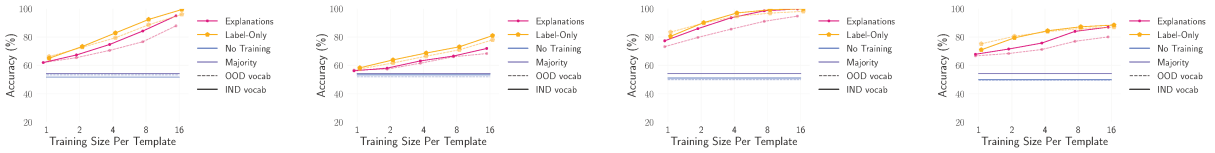
While the BLEU scores can be quite high, OOD generalization remains a challenge. Unseen vocabulary and templates (Fig. 1b, Fig. 1d) increase the difficulty in explanation generation. That said, pretraining on E-SNLI improves generation quality for both IND and OOD cases. This improvement on OOD generalization is likely due to exposure to other data during pretraining.

We use BLEU to evaluate the quality of generated explanations with regard to groundtruth explanations because it is a commonly used metric to evaluate natural language explanations (Camburu et al., 2018; Rajani et al., 2019).

4.2 Predictive Performance

Despite the high BLEU scores, learning with the generated explanations does not help the classification task (Fig. 2). Learning with explanations consistently performs worse than the label-only baseline under both IND and OOD testing scenarios. Pretraining on E-SNLI does not change this observation either.

The only positive result we find is that pretraining helps with OOD generalization. Models pretrained on E-SNLI give better results than plain



(a) IND templates with BERT (b) OOD templates with BERT (c) IND templates with ESNLI (d) OOD templates with ESNLI

Figure 2: x-axis shows the number of samples per template, while $y-$axis shows the accuracy in label prediction. Learning from explanations is always below the label-only baseline.

BERT (Fig. 2). This finding aligns with the positive effect of pretraining on explanation generation.

We also observe that testing on groundtruth explanations boosts performance drastically. This suggests that groundtruth explanations give clues for the label, but generated explanations do not capture this information.

4.3 Why Explanations Do Not Help?

To understand why explanations are not helpful, we introduce two new metrics to evaluate the effectiveness of explanation generation. We measure how often the generated explanations contain **hallucinated entities**, professions (i.e., people) and locations that do not show up in the input, and we measure how well the good label indicator word **"we do not know"** is generated. We present results on explanations generated by the BERT model and the E-SNLI-pretrained model.

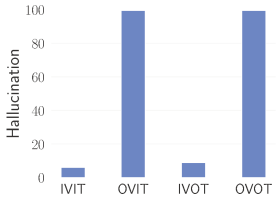
An explanation contains a hallucinated entity if there is an entity that never show up in the original input. These hallucinated entities will likely hinder predictive performance when models make predictions based on generated explanations. We only count hallucinated professions and locations to avoid false positives due to synonyms used in explanations. That is, we use a conservative estimate on hallucinated keywords in generated explanations by only counting people and locations. We find that hallucinated entities are almost always generated in OOD vocab cases by the BERT model (99% of explanations consist of entities that do not exist in the premise and the hypothesis) and the hallucination rate is also high (around 60%) for the E-SNLI-pretrained model. However, the hallucination rate is much lower for IND vocab cases (Fig. 3, Fig. 4): it is close to 0 when $k = 16$ and for E-SNLI-pretrained model. But when $k = 4$, we observe a high hallucination rate ($> 50\%$) for IND vocab cases (Fig. 4a). We also notice that pretraining on E-SNLI leads to models with much lower hallucination rates for all test cases.

IND vocab, IND template ($k = 4$) **Premise**: the managers who the baker addressed brought the technician. **Hypothesis**: the baker addressed the managers. **Original explanations**: who in who the baker addressed refers to the managers. **BERT explanations**: the artisans that addressed the baker are still the managers.
IND vocab, IND template ($k = 16$) **Premise**: the analysts in front of the programmers affected the scientist. **Hypothesis**: the analysts affected the scientist. **Original explanations**: the analysts in front of the programmers are still the analysts. **BERT explanations**: the analysts in front of the programmers are still the analysts.
OOD vocab, IND template ($k = 16$) **Premise**: the chaplains near the singer needed the author. **Hypothesis**: the chaplains needed the author. **Original explanations**: the chaplains near the singer are still the chaplains. **BERT explanations**: the psychologists are in front of the musician and the strategists helped the writer, we do not know whether the illustrators helped the writer.

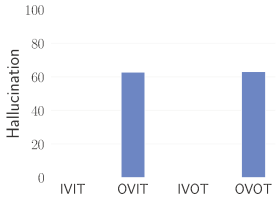
Table 2: Example generated explanations for IND templates cases by the BERT model trained with $k = 4, 16$. More examples are in Appendix C.

"We do not know" is a predictive phrase because it is only present in non-entailment examples. We find that when generated explanations contain "we do not know", so do the corresponding groundtruth explanations (in other words, precision is 100%). However, when "we do not know" is in the groundtruth explanations, it is not necessarily always generated, so the recall is not perfect (Fig. 3c). In fact, recall decreases as we switch to harder test cases. OOD templates also has greater negative impact than OOD vocab on recall.

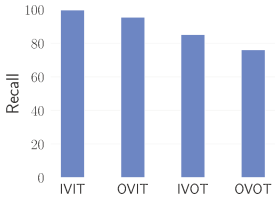
Finally, we look closely at some of the generated explanations (Table 2). We observe that models struggle to learn the templates even for the IND templates case. In the easiest case (*IND vocab, IND template*), although the explanation uses the right template when $k = 16$, it uses a wrong template when $k = 4$. Once we switch from *IND vocab,*



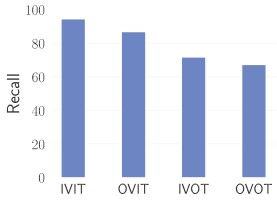
(a) % of explanations with hallucinated entities. (BERT)



(b) % of explanations with hallucinated entities. (E-SNLI)

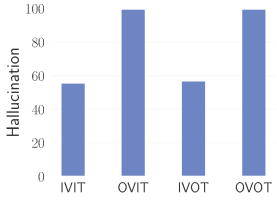


(c) Recall of generating "we do not know". (BERT)

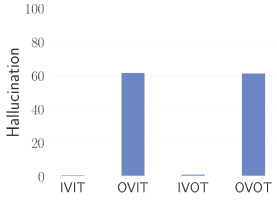


(d) Recall of generating "we do not know". (E-SNLI)

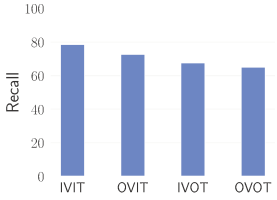
Figure 3: Fig. 3a and Fig. 3b show that the BERT model and the E-SNLI-pretrained model (trained with $k = 16$) hallucinate for OOD vocab. Fig. 3c and Fig. 3d suggest that the explanations fail to include "we do not know" for instances with the non-entailment label for OOD vocab and OOD templates (with $k = 16$).



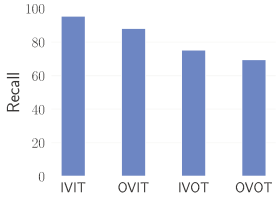
(a) % of explanations with hallucinated entities. (BERT)



(b) % of explanations with hallucinated entities. (E-SNLI)



(c) Recall of generating "we do not know". (BERT)



(d) Recall of generating "we do not know". (E-SNLI)

Figure 4: Fig. 4a and Fig. 4b show that both the BERT model and E-SNLI-pretrained model (trained with $k = 4$) hallucinate for OOD vocab, and the hallucination rate is slightly worse for OOD templates. Similarly, Fig. 4c and Fig. 4d suggest that the explanations fail to include "we do not know" for instances with the non-entailment label for OOD vocab and OOD templates (with $k = 4$).

IND template to *OOD vocab, IND template*, even the $k = 16$ models fail to learn which template should be used to generate explanations.

5 Conclusion

We construct a HANS-based dataset with explanations. On this dataset, we find natural language explanations do not help few-shot NLI to generate out-of-domain under an EXPLAINTHENPREDICT framework. While the genearted explanations obtain high BLEU scores, they do not learn information crucial for downstream classification. Our generation model is close to the SoTA model, yet it still generates nonsensical explanations. Better metrics for explanation evaluation and explanation generation models are key to success for learning with natural language explanations to be effective.

Acknowledgments

We thank anonymous reviewers for their valuable feedbacks. We thank members of the Chicago Human+AI Lab for their insightful suggestions. We thank Tom Mccoy, one author for the HANS paper, for a detailed explanation on their data when we reached out. We thank techstaff members at the University of Chicago CS department for their technical support. This work is supported in part by research awards from Amazon, IBM, Salesforce, and NSF IIS-2126602.

References

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90.

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*, pages 9539–9549.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*.

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

Katharina Kann, Kyunghyun Cho, and Samuel R Bowman. 2019. Towards realistic practices in low-

resource natural language processing: the development set. In *EMNLP*.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *International Conference on Machine Learning*, pages 1078–1088. PMLR.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.

Jonathan Pilault, Amine Elhattami, and Christopher Pal. 2020. Conditionally adaptive multi-task learning: Improving transfer learning in nlp using fewer parameters & less data. *arXiv preprint arXiv:2009.09139*.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*.

Dominik Stammbach and Elliott Ash. 2020. e-fever: Explanations and summaries for automated fact checking. In *Proceedings of the 2020 Truth and Trust Online Conference (TTO 2020)*, page 32. Hacks Hackers.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *arXiv preprint arXiv:2009.10795*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable nlp. ArXiv:2102.12060.

A Replicability Details

We pretrain the BERT model on E-SNLI for 5 epochs and evaluate at every epoch. The model with best dev performance is saved as the final E-SNLI model that we use as the initial model in few-shot learning.

On the E-HANS dataset, we run 2000 steps to train the generation model and evaluate every 200 steps. We choose this number because the best dev performance is usually achieved within 2000 steps. As for the explain-then-predict classifier, we run 200 training steps and evaluate every 4 steps because the model quickly reaches best dev performance as training starts. On the other hand, label-only classifier takes more steps in learning. We train for 1000 steps and evaluate every 50 steps.

It takes around 30 minutes to train a generation model and 10 minutes to train a classification model on our machine (with 4 gpus).

B Samples from E-HANS

Premise: the scientists that talked thanked the psychotherapists. **Hypothesis**: the scientists thanked the psychotherapists. **Explanation**: the scientists that talked are still the scientists.
Premise: if the psychologists ran, the programmers existed. **Hypothesis**: the psychologists ran. **Explanation**: the programmers existed if the psychologists ran, we do not know whether the psychologists ran.
Premise: though the president voted, the musician existed. **Hypothesis**: the president voted. **Explanation**: though suggests the president voted happened.
Premise: the scientist was addressed by the psychotherapist. **Hypothesis**: the psychotherapist addressed the scientist. **Explanation**: addressed is the active form of was addressed by, so we swap the scientist and the psychotherapist.
Premise: if the directors addressed the illustrators, the president listened. **Hypothesis**: the directors addressed the illustrators. **Explanation**: the president listened if the directors addressed the illustrators, we do not know whether the directors addressed the illustrators.

Table 3: Examples from E-HANS.

C Example Generated Explanations

Examples of generated explanations on different test types.

IND vocab, IND template **Premise**: the musician by the psychiatrists offended the strategists. **Hypothesis**: the musician offended the strategists. **Original explanations**: the musician by the psychiatrists is still the musician. **BERT explanations**: the strategists offended the musician and the strategists are by the musician, we do not know whether the strategists offended the musician.
OOD vocab, IND template **Premise**: the administrators near the penciller supported the lyricist. **Hypothesis**: the administrators supported the lyricist. **Original explanations**: the administrators near the penciller are still the administrators. **BERT explanations**: the musician is the subject of the relative clause that addressed the illustrators and the programmers encouraged the programmer, we do not know whether the planners encouraged the illustrators.
IND vocab, OOD template **Premise**: the scientists who affected the colorists helped the psychotherapists. **Hypothesis**: the scientists helped the psychotherapists. **Original explanations**: the scientists who affected the colorists are still the scientists. **BERT explanations**: the scientists are the subject of the relative clause who addressed the psychotherapists and the scientists helped the colorists, we do not know whether the scientists helped the psychotherapists.
OOD vocab, OOD template **Premise**: the professor who deceived the athletes called the doctors. **Hypothesis**: the professor called the doctors. **Original explanations**: the professor who deceived the athletes is still the professor. **BERT explanations**: the psychiatrists who the president offended is still the psychiatrist.

Table 4: Example explanations for all test case generated by the BERT model trained with $k = 4$.

IND vocab, IND template **Premise**: the directors in front of the analysts avoided the designer. **Hypothesis**: the directors avoided the designer. **Original explanations**: the directors in front of the analysts are still the directors. **BERT explanations**: the directors in front of the analysts are still the directors.
OOD vocab, IND template **Premise**: the chaplains near the singer needed the author. **Hypothesis**: the chaplains needed the author. **Original explanations**: the chaplains near the singer are still the chaplains. **BERT explanations**: the psychologists are in front of the musician and the strategists helped the writer, we do not know whether the illustrators helped the writer.
IND vocab, OOD template **Premise**: the technician that thanked the planners encouraged the worker. **Hypothesis**: the technician encouraged the worker. **Original explanations**: the technician that thanked the planners is still the technician. **BERT explanations**: the technician that encouraged the worker is still the technician, so the technician encouraged the worker and we do not know whether the technician encouraged the worker.
OOD vocab, OOD template **Premise**: the senators that recognized the nurses recommended the chaplains. **Hypothesis**: the senators recommended the chaplains. **Original explanations**: the senators that recognized the nurses are still the senators. **BERT explanations**: the presidents are the subject of the relative clause that offended the psychiatrists and the presidents advised the stylists, we do not know whether the officers advised the bakers.

Table 5: Example explanations for all test case generated by the BERT model trained with $k = 16$.