A survey on improving NLP models with human explanations

Mareike Hartmann[1] **Daniel Sonntag**[1,2]
[1]German Research Center for Artificial Intelligence (DFKI), Germany
[2]Applied Artifical Intelligence (AAI), Oldenburg University, Germany
{mareike.hartmann, daniel.sonntag}@dfki.de

Abstract

Training a model with access to human explanations can improve data efficiency and model performance on in- and out-of-domain data. Adding to these empirical findings, similarity with the process of human learning makes learning from explanations a promising way to establish a fruitful human-machine interaction. Several methods have been proposed for improving natural language processing (NLP) models with human explanations, that rely on different explanation types and mechanism for integrating these explanations into the learning process. These methods are rarely compared with each other, making it hard for practitioners to choose the best combination of explanation type and integration mechanism for a specific use-case. In this paper, we give an overview of different methods for learning from human explanations, and discuss different factors that can inform the decision of which method to choose for a specific use-case.

	E-SNLI
Premise	A 2-3 year old **blond** child is kneeling on a couch.
Hypothesis	The child has **brown** hair.
Gold label	Contradiction
Free-text	The child would not have brown hair if he/she was blond.
	COS-E
Questions	What would not be true about a basketball if it **had a hole** in it but it did not lose its general shape?
Answer options	a) punctured, b) full of air, c) round
Gold label	b)
Free-text	Air cannot stay in any object that has a hole in it.

Table 1: Examples of highlight (words marked in bold) and free-text explanations in the E-SNLI dataset (Camburu et al., 2018) for natural language inference and COS-E dataset (Rajani et al., 2019) for multiple choice question answering.

1 Introduction

Enabling machine learning models to learn from human explanations is considered a promising way for interaction between human and machine that can lead to better models and happier users. If a model is provided with information about *why* a specific prediction should be made for an instance, it can often learn more and faster than if just given the correct label assignment (Godbole et al., 2004; Zaidan et al., 2007). This reduces the need for annotated data and makes learning from explanations attractive for use-cases with little annotated data available, for example for adapting models to new domains (Yao et al., 2021) or for personalizing them (Kulesza et al., 2015). Human explanations also push models to focus on relevant features of the data, preventing them from fitting to spurious correlations in the data (Teso and Kersting, 2019). On top of these beneficial effects on model quality, supervision in the form of explanations is in line with human preferences, as users asked to give feedback to a model want to provide richer feedback than just correct labels (Stumpf et al., 2007; Amershi et al., 2014; Ghai et al., 2021).

Several approaches for learning from human explanations have been proposed for different tasks (Table 2), relying on different types of explanations (Table 1), and different methods for integrating them into the learning process. In this paper, we review the literature on learning from highlight and free-text explanations for NLP models, listing technical possibilities and identifying and describing factors that can inform the decision for an optimal learning approach that should optimize both model quality and user satisfaction. Our categorization of methods for integrating explanation information (§ 2.1) is similar to the one provided by Hase and Bansal (2021).[1] Whereas their categorization fo-

[1]Their survey of methods has a broader scope than ours and includes works that improve e.g. image processing models, whereas we exclusively focus on improving NLP models.

cuses on contrasting the approaches according to the role of explanation data in the learning process, we focus on how different types of explanations can be integrated with these approaches.

2 Learning from Explanations

Highlight and free-text explanations are the most prominent explanation types for teaching NLP models (Wiegreffe and Marasovic, 2021). *Highlight explanations* (HIGHLIGHT) are subsets of input elements that are deemed relevant for a prediction.[2] For text-based NLP tasks, they correspond to sets of words, phrases or sentences. *Free-text explanations* (FREE-TEXT) are texts in natural language that are not constrained to be grounded in the input elements. Some recent works rely on *semi-structured text explanations* (SEMI-STRUCTURED) (Wiegreffe and Marasovic, 2021), which combine properties of both highlight and free-text explanations. They consist of text in natural language and contain an explicit indication of the input elements that the free-text applies to.[3] If and how much a model can learn from such explanations depends on the amount of information contained in the explanation (§ 2.2), and to what extent this information can be integrated into the learning process (§ 2.1). User satisfaction is affected by the effort required to produce explanations and by the difficulty of the task, that might in turn affect explanation quality (§ 2.3). In the following, we discuss these factors in detail and where possible contrast them with respect to explanation type.

Objectives Approaches for learning from explanations have been evaluated with different objectives in mind, and we introduce the different motivations below and link them with their respective evaluation in Table 2 (RESULTS column). Early works for learning from explanations were motivated by making the learning process more *efficient* (EFFICIENCY). When human explanations are provided, models can learn more from the same amount of examples (Zaidan et al., 2007), and models can learn more from annotations collected in the same amount of time (Wang et al., 2020), i.e. human labor can be used more efficiently. This makes the paradigm useful for use-cases that allow the collection of additional annotations. Information contained in human explanations can make the model generalize better and lead to better predictive performance on *out-of-domain data* (OUT-OF-DOMAIN), which is most relevant if the model has to be applied under a distribution shift without access to additional annotations. Even with large amounts of annotated data available, models can fit to noise or unwanted biases in the data (Sun et al., 2019), leading to potentially harmful outcomes. Providing human explanations can prevent a model from fitting to such spurious correlations and reduce bias (BIAS REDUCTION).[4] More recently, human explanations have been used in order to improve model explanations (MODEL EXPLANATION, Strout et al. (2019)) or as targets to enable models to generate explanations in the first place (Wiegreffe et al., 2021).

2.1 Integrating explanation information

We now give an overview of different methods[5] that are most commonly applied for integrating the information contained in the human explanation into the model (METHOD column in Table 2).

Given an input sequence $\mathbf{x} = (x_1, \cdots, x_L)$ of length L, a highlight explanation $\mathbf{a}$ is a sequence of attribution scores $\mathbf{a} = (a_1, \cdots, a_L)$, which is of the same length as $\mathbf{x}$ and assigns an importance of $a_i \in \mathbb{R}$ to input element x_i. In practice, a_i is often binary. A free-text explanation $\mathbf{e} = (e_1, \cdots, e_M)$ is a sequence of words of arbitrary length.

Regularizing feature importance This is the dominant approach for learning from highlight explanations. The model is trained by minimizing an augmented loss function $\mathcal{L} = \mathcal{L}_{\text{CLS}} + \mathcal{L}_{\text{EXP}}$ composed of the standard cross-entropy classification loss $\mathcal{L}_{\text{CLS}}$ and an additional explanation loss $\mathcal{L}_{\text{EXP}}$. Given a sequence $\hat{\mathbf{a}} = (\hat{a}_1, \cdots, \hat{a}_L)$ of attribution scores extracted from the model, the explanation loss is computed by measuring the distance between gold attributions a_i and model attribution $\hat{a}_i$ according to $\mathcal{L}_{\text{EXP}}(\mathbf{a}, \hat{\mathbf{a}}) = \sum_{i}^{L} \text{dist}(a_i, \hat{a}_i)$.

$\hat{\mathbf{a}}$ can be extracted from the model using gradient-based or perturbation-based attribution methods (Atanasova et al., 2020). Intuitively, the model

[2] We follow Wiegreffe and Marasovic (2021); Jacovi and Goldberg (2021) in referring to them as highlight explanations.

[3] An overview over NLP datasets with human explanations is provided in Wiegreffe and Marasovic (2021).

[4] For this objective, human explanations are often used as feedback in the *explanation-based debugging* setup, where a bug is identified based on a model's explanation for its prediction and fixed by correcting the model explanation (Lertvittayakumjorn and Toni, 2021).

[5] Hase and Bansal (2021) derive a framework in which some of these methods can be considered as equal.

	TASK								MODEL						EX. TYPE			METHOD					RESULTS			
Authors	DOC. CLS.	SENTIMENT ANALYSIS	HATE SPEECH	REL. EXTRACTION	MULTI CHOICE QA	NATURAL LANG. INFERENCE	EXTRACTIVE QA	OTHER	LOGREG	NAIVE BAYES	SVM	CNN	LSTM	PRETR. TRANSFORMER	HIGHLIGHT	FREE-TEXT	SEMI-STRUCTURED	REGULARIZATION	DATA AUGMENTATION	MTL	I→EX;EX→O	OTHER	OUT-OF-DOMAIN	EFFICIENCY	BIAS REDUCTION	MODEL EXPLANATION
Godbole et al. (2004)	■										■				■							■		▲		
Zaidan et al. (2007)		■									■				■							■		▲		
Zaidan and Eisner (2008)		■													■							■		▲		
Druck et al. (2009)								■							■							■		▲		
Small et al. (2011)		■									■				■							■		▲		
Settles (2011)	■														■							■		▲		
Kulesza et al. (2015)	■									■					■							■		▲		
Zhang et al. (2016)	■											■			■						■			▲		▲
Bao et al. (2018)		■												■	■							■	▲			
Zhong et al. (2019)		■											■		■			■						▲		•
Liu and Avci (2019)			■									■			■			■						▲	▲	
Strout et al. (2019)		■										■			■			■						▲		▲
Rieger et al. (2020)		■											■		■			■							▲	
Stacey et al. (2022)						■								■	■			■					▲	▲		
Carton et al. (2021)	■				■	■								■	■						■			▲		▲
Mathew et al. (2021)			■												■			■							▲	•
Antognini et al. (2021)								■						□	■					■				▲		▲
Pruthi et al. (2022)		■					■							■	■			■		■				▲		
Srivastava et al. (2017)	■								■	■						■			■				▲	▲		
Hancock et al. (2018)				■					■							■			■					▲		
Wang et al. (2020)		■		■									■			■			■					▲		
Lee et al. (2020)		■		■				■					■		■	■			■					▲		
Ye et al. (2020)							■							■			■		■					▲		
Murty et al. (2020)				■										■		■						■		▲		
Yao et al. (2021)		■	■											■			■		■				▲	▲	▲	
Camburu et al. (2018)						■							■			■				■	■		•	•		▲
Rajani et al. (2019)					■									■		■					■		•	▲		▲
Kumar and Talukdar (2020)						■								■		■					■		▼	▲		▲
Zhao and Vydiswaran (2021)						■								■	■	■					■				•	

Table 2: An overview over methods for learning NLP tasks from highlight (upper part) and free-text explanations (lower part). The target task (TASK), model (MODEL), explanation type (EX. TYPE), and integration mechanism (METHOD) used in the respective work is indicated as ■. □ indicates a transformer model without pre-training. For results reported in the respective paper (RESULTS), we explicitly mark an observed increase (▲), decrease (▼), or minimal change (<1%, •) in the evaluated quantity compared to a baseline without access to explanations.

is forced to pay attention to input elements that are highlighted in the highlight explanation. This method is particularly suited for explanatory debugging, as a user can directly interact with a model by modifying the highlight explanations provided by the model.

Semantic parsing to obtain noisy labels This is the dominant approach for learning from free-text explanations. The information contained in the free-text explanations is made accessible via a semantic parser that maps $\mathbf{e}$ to one or more labeling functions λ_i: $\mathcal{X} \rightarrow \{0, 1\}$ (Ratner et al., 2016). λ_i is a logical expression executable on input sequence $\mathbf{x}$ and evaluates to 1 if $\mathbf{e}$ applies to $\mathbf{x}$, and 0 otherwise. The set of all labeling functions is then used to assign noisy labels to unlabeled sequences for augmenting the training dataset. Existing methods differ in how the labeling functions are applied to assign noisy labels, e.g. by aggregating scores over multiple outputs or fuzzily matching input sequences. The approach hinges on the availability of a semantic parser, but several works suggest that using a relatively simple to adapt rule-based parser is sufficient for obtaining decent results (Hancock et al., 2018). Table 2 refers to this approach as DATA AUGMENTATION.

Multi-task learning In the multi-task learning (MTL) approach (Caruana, 1997), two models $\mathrm{M}_{\mathrm{CLS}}$ and $\mathrm{M}_{\mathrm{EXP}}$ are trained simultaneously, one for solving the target task and one for producing explanations, with most of their parameters being shared between them. When learning from high-

light explanations, M_{EXP} is a token-level classifier trained to solve a sequence labeling task to predict the sequence of attributions $\mathbf{a}$. For learning from free-text explanations, M_{EXP} is a language generation model trained to generate the $\mathbf{e}$.

Explain and predict This method was introduced explicitly to improve interpretability of the model, rather than learning from human explanations to improve the target task (Lei et al., 2016). The idea is to first have the model produce an explanation based on the input instance (I→EX), and then predict the output from the explanation alone (EX→O), which is meant to assure that the generated explanation is relevant to the model prediction. The approach can be used for both learning from highlight and free-text explanations.[6] In contrast to the other methods described previously, explain and predict pipelines require explanations at test time. The human explanations are used to train the I→EX component, which provides the EX→O component with model explanations at test time.

Comparative studies We found almost no works that empirically compare approaches for learning from explanations across integration methods or explanation types. Pruthi et al. (2022) compare MTL and REGULARIZATION methods for learning from HIGHLIGHT explanations. They find that the former method requires more training examples and slightly underperforms regularization. Stacey et al. (2022) evaluate their REGULARIZATION method for both HIGHLIGHT and FREE-TEXT explanations. Results are similar for both explanation types, which might be due to the fact that explanations are from the E-SNLI dataset, where annotators were encouraged to include words contained in the highlight explanation into their free-text explanations.

2.2 Information content

Besides the choice of method for integrating explanation information, another important factor affecting model performance relates to the information contained in the explanation. Ideally, we could define specific criteria that determine if an explanation is useful for solving a task, and use these criteria for selecting or generating the most beneficial explanations, e.g. as part of annotation guidelines for collecting explanation annotations. In the following, we summarize findings of recent works that provide insights for identifying such criteria.

Selecting informative explanations Based on experiments with an artificial dataset, Hase and Bansal (2021) conclude that a model can be improved based on explanations if it can infer relevant latent information better from input instance and explanation combined, than from the input instance alone. This property could be quantified according to the metric suggested by Pruthi et al. (2022), who quantify explanation quality as the performance difference between a model trained on input instances and trained with additional explanation annotations. Carton et al. (2021) find that models can profit from those highlight explanations which lead to accurate model predictions if presented to the model in isolation. Carton et al. (2020) evaluate human highlight explanations with respect to their comprehensiveness and sufficiency, two metrics usually applied to evaluate the quality of model explanations (Yu et al., 2019), and observe that it is possible to improve model performance with 'insufficient' highlight explanations. In addition, they find that human explanations do not necessarily fulfill these two criteria, indicating that they are not suited for identifying useful human explanations to learn from. As the criteria listed above depend on a machine learning model, they cannot completely disentangle the effects of information content and how easily this content can be accessed by a model. This issue could be alleviated by using model-independent criteria to categorize information content. For example, Aggarwal et al. (2021) propose to quantify the information contained in a free-text explanation by calculating the number of distinct words (nouns, verbs, adjectives, and adverbs) per explanation.

Explanation type The works described above focus on identifying informative instances of explanations of a given explanation type. On a broader level, the information that can possibly be contained in an explanation is constrained by its type. Highlight explanations cannot carry information beyond the importance of input elements, e.g. world-knowledge relevant to solve the target task, or causal mechanisms (Tan, 2021). Hence, free-text explanations are assumed to be more suitable for tasks requiring complex reasoning, such as natural language inference or commonsense question an-

[6] Wiegreffe et al. (2021) provide a recent survey on explain and predict pipelines. For space reasons, the I→EX;EX→O approaches for learning from HIGHLIGHT explanations listed in their paper are omitted from Table 2.

swering (Wiegreffe and Marasovic, 2021). While this assumption intuitively makes sense, it would be useful to more formally characterize the information conveyed in an explanation of a specific type, in order to match it with the requirements of a given target task. Tan (2021) define a categorization of explanations that might provide a good starting point for characterizing information content. They group explanations into three categories based on the conveyed information: *Proximal mechanisms* convey how to infer a label from the input, *evidence* conveys relevant tokens in the input (and directly maps to highlight explanations), and *procedure* conveys step-by-step rules and is related to task instructions. With respect to matching requirements of a given target task, Jansen et al. (2016) describe a procedure for generating gold explanations covering specific knowledge and inference requirements needed to solve the target task of science exam question answering, which might be transferred to other tasks for generating informative explanations.

2.3 Human factors

Providing explanations instead of just label annotations requires some overhead from the user, which might negatively affect them. Zaidan et al. (2007) found that providing additional highlight explanations took their annotators twice as long as just providing a label for a document classification task.[7] They also point out the necessity to account for human impatience and sloppiness leading to low-quality explanations. Tan (2021) list several factors that might limit the use of human-generated explanations, including their incompleteness and subjectivity. Most importantly, they point out that we cannot expect human explanations to be valid even if the human can assign a correct label, as providing an explanation requires deeper knowledge than label assignment.

3 Take-Aways

While many approaches for enabling NLP models to learn from highlight or free-text explanations have been proposed, there is a lack of comparative studies across different explanation types and integration methods that could reveal the most promising setup to proceed with. Initial studies on the relation between explanation properties and effect on model quality suggest that the explanation's information content plays a central role. We see a promising avenue in developing model-independent measures for quantifying information content, which could be used to give annotators detailed instructions on how to generate an informative explanation that can benefit the model, or to filter out invalid explanations that could harm model performance.

[7] We hypothesize that writing a free-text explanations might take longer than marking highlights for a given task, but could not find any comparison between annotation times for both explanation types.

Acknowledgments

We thank the reviewers for their insightful comments and suggestions. The research was funded by the XAINES project (BMBF, 01IW20005).

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New Dataset and Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, Online. Association for Computational Linguistics.

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120.

Diego Antognini, Claudiu Musat, and Boi Faltings. 2021. Interacting with explanations through critiquing. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 515–521. International Joint Conferences on Artificial Intelligence Organization.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. 2018. Deriving machine attention from human rationales. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1903–1913, Brussels, Belgium. Association for Computational Linguistics.

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, pages 9539–9549. Curran Associates, Inc.

Samuel Carton, Surya Kanoria, and Chenhao Tan. 2021. What to learn, and how: Toward effective learning from rationales. *arXiv preprint arXiv:2112.00071*.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and characterizing human rationales. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9294–9307, Online. Association for Computational Linguistics.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Gregory Druck, Burr Settles, and Andrew McCallum. 2009. Active learning by labeling features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 81–90.

Bhavya Ghai, Q. Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable active learning (XAL): Toward AI explanations as interfaces for machine teachers. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW3).

Shantanu Godbole, Abhay Harpale, Sunita Sarawagi, and Soumen Chakrabarti. 2004. Document classification through interactive supervision of document and term labels. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 185–196. Springer.

Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. Training classifiers with natural language explanations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1884–1895, Melbourne, Australia. Association for Computational Linguistics.

Peter Hase and Mohit Bansal. 2021. When can models learn from explanations? A formal framework for understanding the roles of explanation data. *arXiv preprint arXiv:2102.02201*.

Alon Jacovi and Yoav Goldberg. 2021. Aligning Faithful Interpretations with their Social Attribution. *Transactions of the Association for Computational Linguistics*, 9:294–310.

Peter Jansen, Niranjan Balasubramanian, Mihai Surdeanu, and Peter Clark. 2016. What's in an explanation? characterizing knowledge and inference requirements for elementary science exams. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2956–2965, Osaka, Japan. The COLING 2016 Organizing Committee.

Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, page 126–137, New York, NY, USA. Association for Computing Machinery.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Dong-Ho Lee, Rahul Khanna, Bill Yuchen Lin, Seyeon Lee, Qinyuan Ye, Elizabeth Boschee, Leonardo Neves, and Xiang Ren. 2020. LEAN-LIFE: A label-efficient annotation framework towards learning from explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 372–379, Online. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-Based Human Debugging of NLP Models: A Survey. *Transactions of the Association for Computational Linguistics*, 9:1508–1528.

Frederick Liu and Besim Avci. 2019. Incorporating priors with feature attribution on text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6274–6283, Florence, Italy. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Shikhar Murty, Pang Wei Koh, and Percy Liang. 2020. ExpBERT: Representation engineering with natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2106–2113, Online. Association for Computational Linguistics.

Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C. Lipton, Graham Neubig, and William W. Cohen. 2022. Evaluating Explanations: How Much Do Explanations from the Teacher Aid Students? *Transactions of the Association for Computational Linguistics*, 10:359–375.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. 2020. Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8116–8126. PMLR.

Burr Settles. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Kevin Small, Byron C. Wallace, Carla E. Brodley, and Thomas A. Trikalinos. 2011. The constrained weight space SVM: Learning with ranked features. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 865–872, Madison, WI, USA. Omnipress.

Shashank Srivastava, Igor Labutov, and Tom Mitchell. 2017. Joint concept learning and semantic parsing from natural language explanations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1527–1536, Copenhagen, Denmark. Association for Computational Linguistics.

Joe Stacey, Yonatan Belinkov, and Marek Rei. 2022. Natural language inference with a human touch: Using human explanations to guide model attention. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI 2022)*.

Julia Strout, Ye Zhang, and Raymond Mooney. 2019. Do human rationales improve machine explanations? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–62, Florence, Italy. Association for Computational Linguistics.

Simone Stumpf, Vidya Rajaram, Lida Li, Margaret Burnett, Thomas Dietterich, Erin Sullivan, Russell Drummond, and Jonathan Herlocker. 2007. Toward harnessing user feedback for machine learning. In *Proceedings of the 12th International Conference on Intelligent User Interfaces*, IUI '07, page 82–91, New York, NY, USA. Association for Computing Machinery.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Chenhao Tan. 2021. On the diversity and limits of human explanations. *arXiv preprint arXiv:2106.11988*.

Stefano Teso and Kristian Kersting. 2019. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 239–245, New York, NY, USA. Association for Computing Machinery.

Ziqi Wang, Yujia Qin, Wenxuan Zhou, Jun Yan, Qinyuan Ye, Leonardo Neves, Zhiyuan Liu, and Xiang Ren. 2020. Learning from explanations with neural execution tree. In *International Conference on Learning Representations*.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Huihan Yao, Ying Chen, Qinyuan Ye, Xisen Jin, and Xiang Ren. 2021. Refining language models with compositional explanations. In *Advances in Neural Information Processing Systems*, volume 34, pages 8954–8967. Curran Associates, Inc.

Qinyuan Ye, Xiao Huang, Elizabeth Boschee, and Xiang Ren. 2020. Teaching machine comprehension with compositional explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1599–1615, Online. Association for Computational Linguistics.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4094–4103, Hong Kong, China. Association for Computational Linguistics.

Omar Zaidan and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 31–40, Honolulu, Hawaii. Association for Computational Linguistics.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language*

Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pages 260–267, Rochester, New York. Association for Computational Linguistics.

Ye Zhang, Iain Marshall, and Byron C. Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 795–804, Austin, Texas. Association for Computational Linguistics.

Xinyan Zhao and V.G.Vinod Vydiswaran. 2021. Lirex: Augmenting language inference with relevant explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14532–14539.

Ruiqi Zhong, Steven Shao, and Kathleen McKeown. 2019. Fine-grained sentiment analysis with faithful attention. *arXiv preprint arXiv:1908.06870*.