Explaining the Road Not Taken

Hua Shen
The Pennsylvania State
University
University Park, PA 16802, USA
huashen218@psu.edu

Ting-Hao (Kenneth) Huang
The Pennsylvania State
University
University Park, PA 16802, USA
txh710@psu.edu

ACM CHI Workshop on Operationalizing Human-Centered Perspectives in Explainable AI,, May 8–9, 2021, Online Virtual Conference (originally Yokohama, Japan)
ACM .

Abstract

It is unclear if existing interpretations of deep neural network models respond effectively to the needs of users. This paper summarizes the common *forms* of explanations (such as feature attribution, decision rules, or probes) used in over 200 recent papers about natural language processing (NLP), and compares them against user questions collected in the XAI Question Bank [28]. We found that although users are interested in explanations for *the road not taken* — namely, why the model chose one result and not a well-defined, seemly similar legitimate counterpart — most model interpretations cannot answer these questions.

Introduction

Researchers have attempted to produce model interpretations for deep neural networks [31] under the broader umbrella of Explainable Artificial Intelligence (XAI). The primary objective of this line of research is two-fold [20]: to create interpretations that faithfully characterize the models' behavior (*i.e.*, are *faithful*), and to improve user trust or understanding of black-box algorithms (*i.e.*, appear *plausible*). However, this objective does not always align with the practical needs of users. Recent studies reveal that a faithful or plausible model interpretation can still be useless, or even harmful, to its users. For example, our previous work found that showing users visual explanations (saliency maps) decreased — *not* increased — users' ability to make sense of

Explainable AI Formats-I

1-Feature Attribution (FAT) [43.99%] : highlight the subsequences in input texts [6, 26], Typical question [34]:

- *How can we attribute the systems' predictions to input features?*

2-Tuple/Graph (TUP) [10.15%]: explain model reasoning process with tuples/ trees/ graphs [47, 32]. Typical question [7]:

- *How does the system use reasoning graphs to arrive at the answer?*

3-Concept/Sense (CPT) [9.72%]: convert to human interpretable concepts or terminologies [5, 45]. Typical question [36]:

- *What sense does the system's intermediate representation make?*

4-Rule/Grammar (RUL) [9.61%]: extract executable rules or logic for model decisions. [22, 38]. Typical question [41]:

- *How can we explain the system's behavior with executable rules?*

the mistakes made by neural image classifiers [46]. Another study showed that visual explanations may not alter human accuracy or trust in the model [8]. Recent work in XAI has begun to mitigate this misalignment [9]; one example is collecting algorithm-informed user demands from real-world practices [28].

This paper takes a closer look into the gap between user need and current XAI. Specifically, we survey the common *forms* of explanations, such as feature attribution [6, 26], decision rule [43, 22], or probe [30, 10], used in 218 recent NLP papers, and compare them to the 43 questions collected in the XAI Question Bank [28]. We use the forms of the explanations to gauge the misalignment between user questions and current NLP explanations.

Gauging Explainable AI Gaps Using Forms

Liao *et. al* [28] developed the XAI Question Bank, a set of prototypical questions users might ask about AI systems. This paper investigates how well these questions are answered by current XAI work in NLP. We collected 218 recent NLP papers about interpretability, analyzed the *forms* of interpretations these papers researched (*e.g.*, feature attribution, decision rules, etc.), and used these forms to associate each paper to the questions it tried to answer. This section overviews our two-step procedure.

Step 1: Survey the Forms of Interpretations in NLP Papers. We first reviewed 218 explanation studies published in the NLP field between 2015 and 2020, and came up with 12 common XAI forms. We defined a paper as an NLP explanation study if: *(i)* its motivation was to explain or analyze NLP models, tasks, or datasets; or *(ii)* it aimed to develop more explainable NLP models, tasks, or datasets; or *(iii)* the explanation format is natural language. Given those definitions, we decided on a set of search keywords (*e.g.*, "explain", "interpretation"), a list of top-tier publications and conference proceedings (*e.g.*, ACL, EMNLP), and a range of publication years. Within the venues and years, we collected all papers whose titles or abstracts contained those keywords. Then we read each paper and added the related papers that it cited about "NLP explanation" into our collections. Our ultimate list of papers covered various conferences, workshops, and other research fields (*e.g.*, human-computer interaction).

Our definition of "interpretation form" is *how the study represents its explanation results*. In this paper, we present 12 different interpretation forms. We started with four commonly used forms, including "feature attribution [6, 26]," "tuple/graph [32, 47]," "free text [25, 39]," and "example [16, 54]." Then we read each paper, assigned a form to it, and added new forms into our scheme as needed.

We present the 12 forms with their abbreviations, format weight, brief definitions, representative work, and one typical question in sidebars on page 2 to 4. We released our data[1], which contains the list of the 218 NLP explanation papers with each paper's title, year, venue, and form annotations.

We then computed what percent of the 218 XAI papers used each type of interpretation form (*i.e.,* format weight). We gave each paper a weight of 1. If the paper used only one form type, we assigned 1 to the form. If the paper used multiple interpretation forms, we assigned all its applicable forms an equal weight totaling 1. To obtain the final percentage of each form type, we added up its scores among all papers and divided by the count of papers.

As shown in the sidebars, the most common form of current

[1] Please see details at https://human-centered-exnlp.github.io.

Explainable AI Formats-II

5-Probing (PRB) [7.79%]: classify representation with specific diagnostic dataset [30, 10]. Typical question [18]:

- *What linguistic properties does the system's representation have?*

6-Free Text (FRT) [7.09%]: use natural language to explain model behavior [25, 39]. Typical question [3]:

- *How can we explain a system's decision using natural language justification?*

7-Example (EXP) [3.86%]: find most responsible training samples as explanations [16, 54]. Typical question [24]:

- *How can we trace the system's prediction back to the training sample(s) most responsible for it?*

8-Projection Space (PSP) [3.82%]: project dense vectors into low-dimensional space [48, 52]. Typical question [1]:

- *How can we project the system's high-dimensional representation to a human-understandable space?*

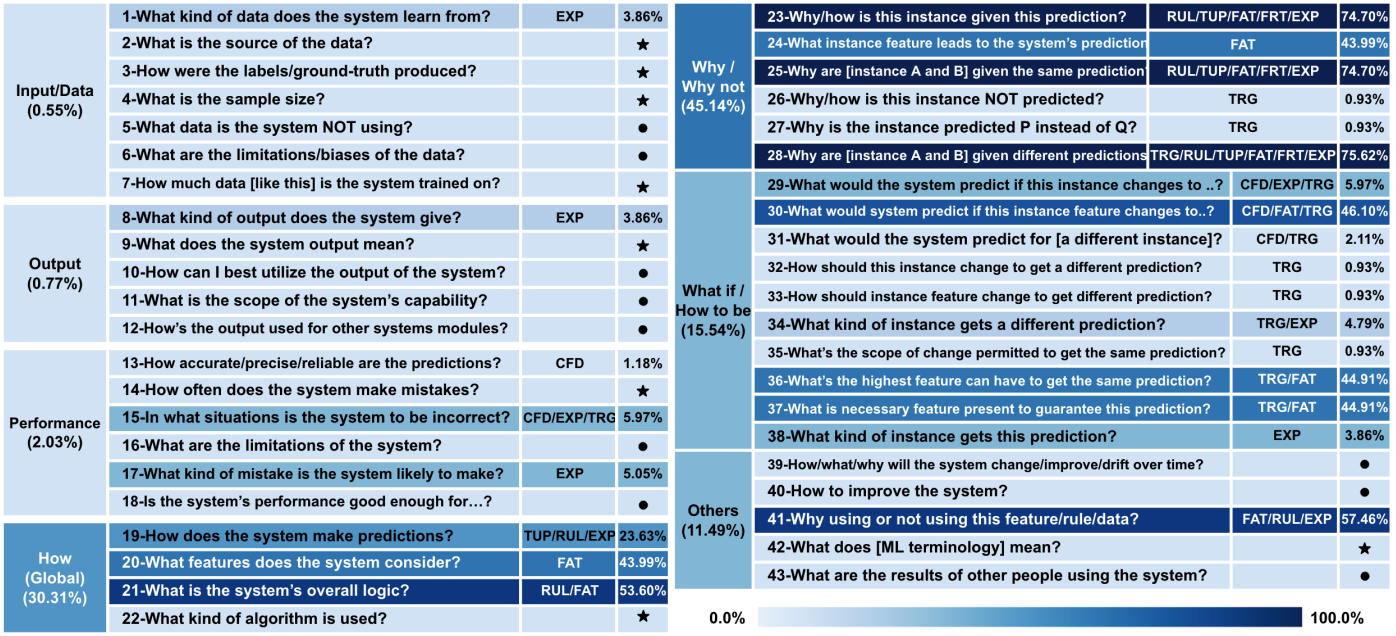

Figure 1: The questions in XAI Question Bank, heat-mapped by the estimated percentage (%) of NLP XAI studies attempting to answer them. (●: questions that can *not* be answered by most NLP XAI studies; ⋆: questions that can likely be answered by the AI system's meta information.)

NLP explanations — around 44% of related studies — is to highlight features (*e.g.*, tokens or sentences) within input text. Approximately 10% of NLP explanation work leverages a tuple, rule, or concept format to demonstrate the model's reasoning process. Other studies use a probe to diagnose what information the model representation can embed, or directly explains model behavior with free text. Less than 5% of algorithms use training data examples, projection space, or output confidence scores to visualize NLP explanations. Fewer than 5 papers explain NLP models with word cloud, trigger, or image formats.

Step 2: Compare Against User Questions in the XAI Question Bank. The XAI Question Bank collected user questions for AI explanations from real-world user needs [28]. It consists of 43 questions within 7 categories about AI systems, as detailed in Figure 1. The prototypical questions are identified by analyzing current XAI algorithms and interviewing UX and design practitioners in IBM product lines. We annotated each user question in the XAI Question Bank with all applicable forms identified in Step 1. The principle we used to annotate a user question with forms was *if the format had ever answered similar questions among our*

Explainable AI Formats-III

9-Confidence Score (CFD) [1.18%]: leverage model prediction probability to show confidence [17, 19]. Typical question [12]:

- *How much uncertainty does the system have on its prediction?*

10-Word Cloud (WCL) [1.16%]: generate word cloud using model representations [37, 6]. Typical question [27]:

- *What are the input patterns that activate the system prediction?*

11-Trigger (TRG) [0.93%]: make change to trigger models to produce counterfactual predictions [13, 44]. Typical question [51]:

- *What are the token sequences that trigger a model to produce a different prediction?*

12-Images (IMG) [0.70%]: visualize model representations by token-related images [35]. Typical question [49]:

- *How to map the system's language tokens to their related images?*

collected studies. Specifically, in the first step, we noted typical questions the form answered in the literature exemplified in the sidebars. For instance, the "example" form primarily answers "What are the training instances most responsible to support this prediction?" [16, 24] Then we inspected each question in the XAI Question Bank and looked for similar questions we collected for the 12 forms. We labeled the user question with its corresponding form when the form was used to answer similar questions in the literature. For instance, to answer the user question "How does the system make a prediction?" we can explain AI systems to users using executable logic rules (*i.e.*, RUL), decision-reasoning graphs (*i.e.*, TUP), or by showing each class's representative examples (*i.e.*, EXP). Afterwards, we calculated each user question's weight by adding all its labeled formats' weights. The user question weight roughly approximates the proportion of published NLP papers that can answer this question. This resulted in the weighted XAI Question Bank as shown in Figure 1, which provides intuitive visualization of the NLP research's attention to answering user questions. Note that XAI forms may evolve rapidly due to proliferation of XAI studies, but we can extend the collected XAI forms and repeat the gap-gauging process easily.

The Need to Explain the Road Not Taken

While 9 out of 43 questions in the XAI Question Bank are about how AI systems **can** provide specific predictions (*i.e.*, Q19-21,23-25,36-38), 16 questions are about what AI systems **cannot** achieve and why (*e.g.*, Q5-6,11,15-16,26-35,41). Many of these under-answered or unanswered questions are *counterfactual* questions, such as "Why did the model predict P instead of Q for this instance?" These questions can probably be answered by a trigger (TRG), but only three papers out of the surveyed 218 focused on counterfactual explanations [51, 13, 44]. Furthermore, we speculate that many of these questions assume one or more well-defined, seemly similar legitimate counterpart labels (*e.g.*, *positive* versus *negative*, *dog* versus *cat*), in which the user wonders why the system choose one over the other. More fundamentally, the fact that users want to know both *why* and *why not* the AI system made certain predictions may suggest that users' goals are often to **gain a *global view* of how the AI system works**.

It is worth noting that more NLP work has begun to generate counterfactual examples (*i.e.*, "contrastive sets"), often with the purpose of learning robust NLP models [14, 23, 53]. These methods could be extended to generating counterfactual explanations. As counterfactual explanations have been explored in other domains, such as computer vision [4], tabular data classification [33], and interactive tools [15], recent NLP work has begun to focus more attention on developing counterfactual explanations [21, 42][2].

Which Road Do You Want Explanations For? Developing counterfactual explanations in NLP can be challenging. It is not always easy to tell **which counterfactual predictions** should be explained. Jacovi *et al.* submitted a good example [21]: When people ask "Why did the AI system choose to hire Person X?" they could mean either "Why did the AI system choose to hire Person X rather than not hire Person X?" or "Why did the AI system choose to hire Person X rather than hire Person Y?" Liao *et al.* suggested that AI explanations can be provided in an *interactive* manner, allowing people to "explicitly reference the contrastive outcome and ask follow-up *what if* questions" [28]. As ambiguous and underspecified language can be common, more research is required to help users spot the meaningful counterfactual predictions they actually care about.

[2]We did not include these recent studies in our paper collection because they were published after our paper-collecting and analysis process.

Discussion

User Questions Beyond the Scope of the Current XAI. In another finding included in Figure 1, we observed 8 questions (*i.e.*, labeled ⋆) that can be addressed by the *meta information* in AI algorithms (such as "What is the source of the data?") but that XAI forms do not answer. However, we find 10 questions (*i.e.*, labeled •) that the XAI forms cannot address well. These questions mainly inquire about the **limitation, potential utility, or capability scope** of AI systems (*e.g.*, "What are the limitation/biases of the data?"), which are seldom introduced in XAI studies. We posit XAI algorithm developers should use these questions to develop corresponding XAI methods or to clarify capability scope, system utility, and limitation in the methods.

Limitations. We are aware of several limitations of our work. First, this paper focuses on NLP applications, but the XAI Question Bank captures user questions for a broader spectrum of AI systems. Second, the XAI Question Bank provides an in-depth analysis of lay users' needs, while the user population for the NLP papers included in our study are broader, such as domain experts [11, 50] and AI practitioners [40, 2]. Finally, using forms of interpretation to associate papers with user questions inevitably overlooks some information. For instance, the "probing" form does not appear in the XAI Question Bank. This could be caused by the fact that some particular forms of interpretations, such as probing methods, are primarily developed for AI practitioners rather than lay people.

Conclusion

Our analysis explicates the gaps between what users want and the current focus of XAI research in NLP. Questions like "Why is this instance given this prediction?" were studied extensively, and can be answered by five different interpretation formats (*i.e.*, "rule/grammar," "tuple/graph," "feature importance," "free text," and "example"). Meanwhile, 16 out of 43 user questions in the XAI Question Bank are relevant to counterfactual inquiries, such as "Why did the model predict P instead of Q for this instance?", but only a handful of papers have tried to produce counterfactual explanations. We learned that users want to know the decision scope of AI systems, including what the AI system can and cannot achieve.

XAI researchers can collaborate with user-experience (UX) designers to mitigate this misalignment. In particular, XAI algorithm developers can produce more counterfactual explanations for answering global and local counterfactual questions, or directly generate AI explanations that can explain both *can* and *cannot* questions (*e.g.*, tree-based rules). On the other hand, one XAI form may not be enough to satisfy practical user demands for understanding *can* and *cannot* questions simultaneously. Therefore, XAI UX designers can combine multiple forms and algorithms to meet real-world user requirements. Since awareness of new explainable AI forms can change user demand [29, 28], perhaps XAI researchers can leverage the variety of forms to to respond more effectively to real-world user needs.

REFERENCES

[1] Malika Aubakirova and Mohit Bansal. 2016. Interpreting Neural Networks to Improve Politeness Comprehension. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2035–2041. DOI: http://dx.doi.org/10.18653/v1/D16-1216

[2] Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?. In *Proceedings of the Third BlackboxNLP Workshop on*

Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics, Online, 149–155. https://www.aclweb.org/anthology/2020.blackboxnlp-1.14

[3] 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*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc., 9539–9549. https://proceedings.neurips.cc/paper/2018/file/4c7a167bb329bd92580a99ce422d6fa6-Paper.pdf

[4] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. 2019. Explaining Image Classifiers by Counterfactual Generation. In *International Conference on Learning Representations*. https://openreview.net/forum?id=B1MXz20cYQ

[5] Ting-Yun Chang and Yun-Nung Chen. 2019. What does this word mean? explaining contextualized embeddings with natural language definition. 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)*. 6066–6072.

[6] Hanjie Chen and Yangfeng Ji. 2020. Learning Variational Word Masks to Improve the Interpretability of Neural Text Classifiers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 4236–4251. https://www.aclweb.org/anthology/2020.emnlp-main.347

[7] Jifan Chen, Shih-ting Lin, and Greg Durrett. 2019. Multi-hop question answering via reasoning chains. *arXiv preprint arXiv:1910.02610* (2019).

[8] Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248* (2020).

[9] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing transparency design into practice. In *23rd international conference on intelligent user interfaces*. 211–223.

[10] Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics* 8 (2020), 34–48.

[11] Jinyue Feng, Chantal Shaib, and Frank Rudzicz. 2020. Explainable Clinical Decision Support from Text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 1478–1489. https://www.aclweb.org/anthology/2020.emnlp-main.115

[12] Shi Feng and Jordan Boyd-Graber. 2019. What can AI do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 229–239.

[13] Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of Neural Models Make Interpretations Difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 3719–3728.

[14] Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating Models' Local Decision Boundaries via Contrast Sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 1307–1323. https://www.aclweb.org/anthology/2020.findings-emnlp.117

[15] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. 2020. ViCE: Visual Counterfactual Explanations for Machine Learning Models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)*. Association for Computing Machinery, New York, NY, USA, 531–535. DOI:http://dx.doi.org/10.1145/3377325.3377536

[16] Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. 2020. Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions. *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)* (2020).

[17] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? *arXiv preprint arXiv:2005.01831* (2020).

[18] John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2019).

[19] Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2020. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 187–196.

[20] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?. In *ACL'20*. Association for Computational Linguistics, Online, 4198–4205. DOI: http://dx.doi.org/10.18653/v1/2020.acl-main.386

[21] Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive Explanations for Model Interpretability. *arXiv preprint arXiv:2103.01378* (2021).

[22] Yichen Jiang and Mohit Bansal. 2019. Self-Assembling Modular Networks for Interpretable Multi-Hop Reasoning. 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)*. Association for Computational Linguistics, 4474–4484.

[23] Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning The Difference That Makes A Difference With Counterfactually-Augmented Data. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Sklgs0NFvr

[24] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. *Proceedings of IEEE Conference on Machine Learning (ICML)* (2017).

[25] 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*. Association for Computational Linguistics, Online, 372–379. DOI: http://dx.doi.org/10.18653/v1/2020.acl-demos.42

[26] Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 107–117.

[27] Piyawat Lertvittayakumjorn, Lucia Specia, and Francesca Toni. 2020. FIND: human-in-the-loop debugging deep text classifiers. *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2020).

[28] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[29] Brian Y Lim and Anind K Dey. 2009. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing*. 195–204.

[30] Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open Sesame: Getting inside BERT's Linguistic Knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Florence, Italy, 241–253. DOI:http://dx.doi.org/10.18653/v1/W19-4825

[31] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73 (2018), 1–15.

[32] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 845–854.

[33] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 607–617. DOI: http://dx.doi.org/10.1145/3351095.3372850

[34] Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the Model Understand the Question?. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1896–1906.

[35] Alexander Panchenko, Fide Marten, Eugen Ruppert, Stefano Faralli, Dmitry Ustalov, Simone Paolo Ponzetto, and Chris Biemann. 2017. Unsupervised, knowledge-free, and interpretable word sense disambiguation. *arXiv preprint arXiv:1707.06878* (2017).

[36] Abhishek Panigrahi, Harsha Vardhan Simhadri, and Chiranjib Bhattacharyya. 2019. Word2Sense: Sparse Interpretable Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 5692–5705. DOI:http://dx.doi.org/10.18653/v1/P19-1570

[37] Nikolaos Pappas and Andrei Popescu-Belis. 2014. Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis. In *Proceedings of the 2014 Conference on Empirical Methods In Natural Language Processing (EMNLP)*. 455–466.

[38] Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2019. Investigating Robustness and Interpretability of Link Prediction via Adversarial Modifications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 3336–3347.

[39] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)* (2019).

[40] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "" Why should I trust you?"" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[41] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations.. In *AAAI*, Vol. 18. 1527–1535.

[42] Alexis Ross, Ana Marasović, and Matthew E Peters. 2020. Explaining nlp models via minimal contrastive editing (mice). *arXiv preprint arXiv:2012.13985* (2020).

[43] Swarnadeep Saha, Sayan Ghosh, Shashank Srivastava, and Mohit Bansal. 2020. PRover: Proof Generation for Interpretable Reasoning over Rules. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 122–136. https://www.aclweb.org/anthology/2020.emnlp-main.9

[44] Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do Neural Dialog Systems Use the Conversation History Effectively? An Empirical Study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 32–37.

[45] Robert Schwarzenberg, Lisa Raithel, and David Harbecke. 2019. Neural vector conceptualization for word vector space interpretation. *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (2019).

[46] Hua Shen and Ting-Hao Huang. 2020. How Useful Are the Machine-Generated Interpretations to General Users? A Human Evaluation on Guessing the Incorrectly Predicted Labels. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 168–172.

[47] Josua Stadelmaier and Sebastian Padó. 2019. Modeling Paths for Explainable Knowledge Base Completion. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 147–157.

[48] Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander M Rush. 2018. Seq2seq-vis: A visual debugging tool for sequence-to-sequence models. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 353–363.

[49] Hao Tan and Mohit Bansal. 2020. Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2020).

[50] Marco A Valenzuela-Escárcega, Ajay Nagesh, and Mihai Surdeanu. 2019. Lightly-supervised representation learning with global interpretability. *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (2019).

[51] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. *arXiv preprint arXiv:1908.07125* (2019).

[52] Jialin Wu and Raymond Mooney. 2019. Self-Critical Reasoning for Robust Visual Question Answering. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc., 8604–8614.

[53] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Automated, General-purpose Counterfactual Generation. *arXiv preprint arXiv:2101.00288* (2021).

[54] Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. 2018. Representer point selection for explaining deep neural networks. In *Advances in neural information processing systems*. 9291–9301.