

## Book Review

**Explainable Natural Language Processing, by Anders Søgaard. San Rafael, CA: Morgan & Claypool, 2021. ISBN 978-1-636-39213-4. XV+107 pages.**

In the age of deep learning, Richard Feynman’s famous quote ‘what I cannot create, I do not understand’ might tell only half the truth. Even if researchers nowadays can create models that work reasonably well, they have little understanding of how they work that well. However, the issue of model explainability is such an important one which cannot be simply ignored. There are multiple justifications behind our quest for explainable models, including safety reasons, privacy concerns, ethical bias, social acceptance, model improvement and goal of science (Doshi-Velez and Kim 2017; Molnar 2022). In fact, an explainable model with competitive performance is always preferred over a non-explainable one given its transparency (Ribeiro, Singh, and Guestrin 2016). Explainable natural language processing (NLP) is an emerging track of NLP research and has been an independent field of NLP research since 2018 (Alishahi, Chrupała, and Linzen 2019). Since a reasonable classification or taxonomy is always a prerequisite for a fundamental theory for an emerging field, the volume under review contributes to the NLP community by presenting a novel taxonomy of explainable NLP methods, as well as a comprehensive review of them. Although the author claims that the book is readable for both first-year M.Sc. students and expert audience, the brevity of style and the broad coverage of the book probably make it more suitable for readers towards the more experienced end of the range.

The book consists of 13 chapters. Chapter 1 briefly discusses the current problems in explainable NLP and highlights the need for a well-formed taxonomy. A critical and systematic review of the existing taxonomies forms the bulk of this chapter. Based on this, in Chapter 2, following a brief introduction to the current popular NLP architectures, the author outlines a novel two-dimensional taxonomy of explainable NLP methods. From Chapter 3 to Chapter 10, the author explains the proposed taxonomy by briefly surveying each of its classes. Chapter 11 provides guidelines for the evaluation of explanations. In Chapter 12, the author demonstrates the role of the taxonomy as a framework for discovering new possible explanation methods, while Chapter 13 closes the book with a list of useful resources.

Chapter 1 identifies the currently existing problems in the emerging explainable NLP subfield, that is, without a well-formed taxonomy, researchers might work on questions that have already been studied and provide methods that have already been invented. The author thus highlights the importance of a carefully constructed taxonomy of the existing explanation methods as a solution to the problems. In this chapter, the author presents a critical and systematic review of 10 popular taxonomies, which is one of the most important contributions of the book. The author claims that most existing taxonomies have at least one of three major shortcomings, namely incompleteness, inconsistency and redundancy. To expose the shortcomings of previous taxonomies more concretely, the author poses a challenge to each listed taxonomy in turn by showing a classification task which would be hard or impossible to be accomplished using it. Another common shortcoming of the taxonomies is that they pay too much attention to the form of the explanations, which is orthogonal to the aim of building a consistent taxonomy of methods. The author refers to this shortcoming as *method-form fallacy*.

In Chapter 2, after a brief introduction to existing NLP models (from linear models to transformers) along with their applications, the author explains his two-dimensional taxonomy, with one dimension known as ‘local versus global’ and the other as ‘forward versus backward’. ‘An interpretability method is said to be *global* if and only if its explanations rely on access to an (i.i.d.) *sample* of representative instances; otherwise, if the method can provide explanations for individual instances in the absence of such samples, it is said to be *local*’ (p. 3). ‘An interpretability method is said to be *backward* if it relies solely on quantities derived from one or more backward passes through the instances; otherwise, if it relies on quantities from forward passes, it is said to be *forward*’ (p. 12). The author further identifies two types of forward explainability methods (both local and global): those that focus on intermediate versus output representations. Methods which focus on model outputs are further divided by whether the output representations are continuous or discrete. With this division, we end up with eight categories of explainability methods, which are exactly the skeleton of the next eight chapters.

Based on this taxonomy, the author places into the eight groups 32 kinds of existing popular explainability methods. Chapter 3 covers *local-backward explanations*, describing six methods which take advantage of gradients or relevance scores to obtain explanations. Vanilla gradients and guided back-propagation are among the earliest and simplest gradient-based methods here. The common idea of these methods is to compute the gradients of the models’ prediction with respect to the inputs, leaving the model weights fixed. Instead of relying on gradients, layer-wise relevance propagation back-propagates relevance scores recursively to the input layer to get the model explanation. Deep Taylor decomposition applies Taylor decomposition at some well-chosen root point  $\tilde{x}$  to redistribute relevances to lower layers. Similar to the idea of root point, integrated gradients and DeepLIFT use neutral reference points as baselines as well.

Chapter 4, *Global-Backward Explanations*, covers four methods of explainability which assumes that sparse models are inherently more interpretable. Methods that fall into this category prune or sparsify models to achieve the desired simplicity. Sparse coding tries to represent the input feature by the activation of a small set of neurons, while binary networks use binary weights instead of real-valued weights. Dynamic sparse training and lottery tickets are both pruning methods: the latter re-trains the model after pruning, while the former jointly trains and prunes networks.

Chapter 5, *Local-Forward Explanations of Intermediate Representations*, focuses mainly on easily interpretable model components, namely gate activations of recurrent neural networks (RNNs) and attention matrices of transformers. Gate activations are retrieved to visualise the inner workings of recurrent models and to detect the encoded linguistic properties. In a similar spirit, different schemes have been devised to work with attention weights, namely attention roll-out, attention flow, layer-wise attention tracing and attention decoding, to achieve model explainability.

Chapter 6, *Global-Forward Explanations of Intermediate Representations*, also focuses on gate activations and attention weights (i.e., intermediate representations). As global methods, these rely on pruning techniques to increase the understanding of models under investigation. Research on both model components is based on the same assumption that ‘the extent to which Gates or Attention heads can be removed tells us a lot about the inner workings of a neural network’ (p. 35).

Explanation methods in Chapter 7, *Local-forward Explanations of Continuous Output*, make use of word embeddings (as the outputs of models) to understand the models under investigation and include the widely used methods of word association norms and word analogies, as well as the more recently proposed time step dynamics. Word association norms and word analogies are considered classical means of evaluating the quality of word representations. Time step dynamics works with RNN-based models, with explanatory analysis ranging from simple activation plots to more complicated contextual decomposition.

Chapter 8, *Global-Forward Explanations of Continuous Output*, covers five explanation methods. Correlation of representations methods sees the output of a sample of input examples as point clouds, where the correlation between these clouds can be interpreted. Clustering explores how language models encode linguistic concepts through the distributions of the output vectors of models. The idea behind probing classifiers is that, by explicit supervision, if a model can classify (or predict) linguistic properties with high accuracy, it learns the property. The concept activation method originates from the computer vision community. It differs from gradient-based methods in that it attributes importance to high-level concepts instead of low-level features. Although it makes intuitive sense to adapt this method for NLP (given its similarity to probing classifiers), it fails to receive equally wide adoption, as it is hard to perceive concepts of relevance in NLP tasks in a continuous way. Finally, influential examples methods provide an explanation for model decisions in terms of training instances.

Three explanation methods fall into the category detailed in Chapter 9, *Local-Forward Explanations of Discrete Output*. Challenge datasets are carefully designed tests for language models to investigate their ability to handle particular linguistic phenomena. These tests are often backed with solid linguistic theories. Local uptraining methods learn simple approximations of neural networks, and the simple models obtained are easily interpretable. The most representative method among these uptraining techniques is Local Interpretable Model-agnostic Explanations (LIME), which is widely used in the NLP community. Meanwhile, influential examples methods show us that explanations do not necessarily need to be in the form of importance attribution or visualisations. Instead, they find the most influential example in training data that correlates with the discrete outputs of models.


Chapter 10, *Global-forward Explanations of Discrete Output*, involves three explainability methods. The idea behind global uptraining is similar to that of local uptraining in Chapter 9, except that it works with a sample of inputs. Meta-analysis induces a regressor (e.g. a linear or lasso regression) to predict the performance of the model under investigation on different datasets. Downstream evaluations are intuitive methods of model explanation, as they show in a human understandable way what the models are good at and where they fail, thus providing clear guidance on where the models can be improved.

In Chapter 11, *Evaluating Explanations*, the author discusses two topics relating to Explainable NLP, namely, forms and evaluations of explanations. Following DeYoung *et al.* (2019), the author identifies four forms of explanations. Extractive rationales highlight subsets or substrings of the input. Abstractive rationales typically consist of concepts, logical or linguistic structures, or human-readable texts. Explanations in the form of training instances are simply a subset of the training examples. Model visualisations can be conducted on model outputs or intermediate representations to make sense of the components of their inner workings. In the rest of this chapter, following Doshi-Velez and Kim (2017), the author introduces his revised classification of methods of evaluating explanations, namely heuristics, human annotation and human experiments.

The observations provided in Chapter 12, *Perspectives*, extend the work beyond simply categorising methods. With the help of the proposed taxonomy, the author suggests 11 principles for seeking or designing new explainability methods. These inspiring observations make the proposed taxonomy not only an arrangement of methods but also an early form of explainability theory. Finally, Chapter 13, *Resources*, provides the readers with additional code, datasets and benchmarks which are useful for both researchers and NLP practitioners.

Overall, *Explainable Natural Language Processing* provides the NLP community with a clear-cut, consistent and inclusive taxonomy of explainable NLP methods. With the help of the meticulously designed taxonomy, researchers can collaborate under consistent terms, seek explanation methods in principled ways and evaluate results under common standards. Furthermore, works that have not been previously considered as explainable NLP methods (e.g., the pruning methods in Chapter 4) are incorporated into the two-dimensional taxonomy. Although the book covers a broad range of explainable NLP work, it should not simply be seen as a review of

existing methods. Instead, it goes beyond simply grouping these methods, by revealing the correlations between them and providing suggestions for new methods. More importantly, the insightful observations (as in Chapter 12), discussing what is possible and what is not, provide researchers with a vehicle for thinking about potential new methods in a principled way.

Zihao Zhang 

Institute of Quantitative Linguistics, Beijing Language and Culture University, Beijing 100083, China

E-mail: [johnthehow@qq.com](mailto:johnthehow@qq.com)

## References

- Alishahi A., Chrupała G. and Linzen T.** (2019). Analyzing and interpreting neural networks for NLP: a report on the first BlackboxNLP workshop. *Natural Language Engineering* 25(4), 543–557.
- DeYoung J., Jain S., Rajani N.F., Lehman E., Xiong C., Socher R. and Wallace B.C.** (2019). ERASER: a benchmark to evaluate rationalized NLP models. ArXiv Preprint ArXiv:1911.03429.
- Doshi-Velez F. and Kim B.** (2017). Towards a rigorous science of interpretable machine learning. ArXiv Preprint ArXiv:1702.08608.
- Molnar C.** (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd edn. <https://christophm.github.io/interpretable-ml-book>.
- Ribeiro M.T., Singh S. and Guestrin C.** (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, pp. 1135–1144.