Book Review

Explainable Natural Language Processing

Anders Søgaard
(University of Copenhagen)

Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 51), 2021, xvi + 107 pp; paperback, ISBN: 9781636392134; ebook, ISBN: 9781636392141; hardcover, ISBN: 9781636392158
DOI:10.2200/S01118ED1V01Y202107HLT051

Reviewed by
George Chrysostomou
University of Sheffield

Explainable Natural Language Processing (NLP) is an emerging field, which has received significant attention from the NLP community in the last few years. At its core is the need to explain the predictions of machine learning models, now more frequently deployed and used in sensitive areas such as healthcare and law. The rapid developments in the area of explainable NLP have led to somewhat disconnected groups of studies working on these areas. This disconnect results in researchers adopting various definitions for similar problems, while also in certain cases enabling the re-creation of previous research, highlighting the need for a unified framework for explainable NLP.

Written by Anders Søgaard, this book provides the author's convincing view of how we should first define explanations, and, secondly, how we should categorize explanations and the approaches that generate them, creating first and foremost a taxonomy and a unified framework for explainable NLP. As per the author, this will make it easier to relate studies and explanation methodologies in this field, with the aim of accelerating research. It is a brilliant book for both researchers starting to explore explainable NLP, but also for researchers with experience in this area, as it provides a holistic up-to-date view of the explainable NLP at the *local and global level*. The author conveniently and logically presents each chapter as a "problem" of explainable NLP, as such providing also a taxonomy of explainable NLP problem areas and current approaches to tackle them. Under each chapter, explanation methods are described in detail, beginning initially with "foundational" approaches (e.g., vanilla gradients) and building toward more complex ones (e.g., integrated gradients). To complement the theory and make this into a complete guide to explainable NLP, the author also describes evaluation approaches and provides a list of datasets and code repositories. As such, although the book requires some basic knowledge of NLP and Machine Learning to get started, it is nevertheless accessible to a large audience.

This book is organized into thirteen chapters. In the first chapter the author introduces the problems associated with previously proposed taxonomies for explainable NLP. Chapter 2 follows by introducing popular machine learning architectures used in NLP, while also introducing the explanation taxonomy proposed in the book. Chapters

https://doi.org/10.1162/coli_r_00460

3 and 4 describe explanation methodologies for extracting local and global explanations, respectively, that require performing a backward pass through the model. Chapters 5–10 are focused on local and global explanation methods, that require only a forward pass through the model for different types of output: (a) intermediate representations; (b) continuous output; and (c) discrete output. Chapter 11 then describes how we can evaluate these explanation methods, and Chapter 12 provides perspectives on the proposed taxonomy and concludes this work. Finally, Chapter 13 provides resources for explainable NLP.

Chapter 1 introduces explainable NLP and the problems associated with the lack of a unified framework. It then continues to justify the importance of taxonomies, while briefly describing the two high-level categories for the proposed taxonomy: (a) **local** and **global**, (b) **forward** and **backward**. For the former category, a method is considered *local* when it can provide explanations for individual instances, otherwise it is considered *global*. The latter distinguishes between methods that rely on *forward* passes over the parameters and those that rely on *backward* passes. A large part of this chapter is then dedicated to describing previous efforts in creating taxonomies, while also highlighting their problems.

Chapter 2 begins by giving a brief overview of popular machine learning models used in NLP, such as the transformers, followed by other popular NLP applications. It then follows by proposing further low-level categories that operate through the forward pass on (a) intermediate representations; (b) continuous output; and (c) discrete output, providing a clearer structure and categorization to explanation methods. The sections that follow then justify the importance of the previously described high-and-low-level categories, by providing examples of their uses and links to future chapters.

Chapter 3 focuses on local-backward explanation methods that use training signals to explain model predictions, without additional parameters. This includes and describes mathematically popular explanation approaches starting from simple methods, such as vanilla gradients, building up to DeepLift. A pleasant addition is the use of open problems, where the authors, after the description of an explanation method, pose logical questions that make the reader think and that motivate good research directions. The following chapter (Chapter 4) describes popular global-backward explanation methods such as post-hoc unstructured pruning and binary networks. A commendable inclusion is the discussion around computational cost of some of the methods.

Chapter 5 is the first of the chapters focusing on the forward distinction, describing local-forward explanation methods for intermediate representations. A popular explanation method belonging in this category is attention weights, although more recent methods such as attention flow are also described. Under each explanation method, where applicable, the author also includes comments from previous studies on their efficacy. Additionally, the author demonstrates that two methods developed concurrently in literature are the same, reinforcing their initial claim that the lack of taxonomy and a unified framework often leads to duplicated work. Chapter 6 then follows with global-forward explanation methods for intermediate representations, such as attention head pruning.

Describing local-forward explanation methods on continuous output vectors, Chapter 7 highlights three different approaches. As continuous output vectors (e.g., word representations) are used typically in downstream tasks, the author describes three valid explanation methods: word association norms, word analogies, and time-step dynamics. Chapter 8 then describes forward explanation methods that operate at the global-level for continuous outputs, such as probing classifiers, clustering, and finding the most influential training examples.

Chapter 9 continues with forward explanation methods, focusing on those that operate at the local level illustrating methods that can explain a model's decision based on its discrete output. Methods include creating challenge datasets, a set of examples to test the hypothesis, for example, that certain inputs affect the model's predictions, finding influential examples *at the local level*, and uptraining methods such as LIME. Chapter 10 then extends on discrete outputs, describing explanation methods that operate at the global level, such as understanding how good models are using downstream evaluation or using simpler models such as linear regression to obtain sparser, more interpretable explanations.

Chapter 11 introduces the reader to how to evaluate explanations, a widely discussed topic in the explainable NLP research community. The author conveniently discretizes first explanations into four distinct categories: extractive rationales (e.g., heatmaps); abstractive rationales (e.g., concepts); training instances (e.g., examples); and model visualizations. This discretization also continues over to evaluation methods, splitting them into three categories: heuristics; human annotations; and human experiments, where each explanation category is associated with different evaluation method categories. Following these distinctions, the chapter then describes, under each evaluation method category, popular evaluation methods and the types of explanations they cover. Each subsection logically includes the limitations and open-problems associated with each of the evaluation categories, allowing the reader to follow research directions based on these open-problems, while maintaining order within this framework and taxonomy.

Chapter 12 concludes the previous chapters by providing a set of observations gathered from this taxonomy. Following these, it then describes concepts that are not covered by this taxonomy, such as contrastive and comparative explanations. At the end of this chapter the author discusses the ethical and moral foundations of explanations, presenting arguments from literature and critiquing the basic assumption that human decision making is fully transparent, when in fact it is not. The final part of Chapter 12 in particular is extremely interesting for the reader, as it helps put in perspective what we as researchers expect out of explainable NLP when in fact human decision making is non-trivial to explain. Finally, to help researchers delve into explainable NLP and expand their knowledge, the author conveniently provides a list of resources (i.e., data, benchmarks, and code) in Chapter 13.

George Chrysostomou is an NLP Engineer at AstraZeneca, who has recently completed his Ph.D. studies at the University of Sheffield. His main research interests and Ph.D. projects revolve around explainable NLP and particularly interpreting model predictions in NLP applications. Other research interests include improving the efficiency in pre-training large language models and low-resource named entity recognition. George's e-mail address is gchrysostomou1@sheffield.ac.uk.