Explainable Natural Language Processing (NLP)

Atreya Shankar
atreya.shankar@{uni-potsdam.de,uzh.ch}
Department of Linguistics, University of Potsdam
Department of Computational Linguistics, University of Zürich

October 15, 2020

Abstract

Recent advancements in Natural Language Processing (NLP) have been largely driven by developments in Deep Learning techniques. While powerful, these methods pose a serious limitation of being "black-box" or opaque. This has led to problems such as adversarial attacks and inherent biases leading to downstream ethical issues. The objective of this research is to explore white-box and grey-box machine learning methods in NLP which offer explainability, while still providing competitive performance compared to their deep learning counterparts.

Contents

1 Literature review

1.1 Explainable artificial intelligence

Doran et al. 2017: This study conducts a review of Explainable Artificial Intelligence (XAI) based on corpus-level analyses of NIPS, ACL, COGSCI and ICCV/ECCV papers. Based on their analysis, they characterize existing XAI into three mutually exclusive categories; namely *opaque systems* that offer no insight into internal model mechanisms, *interpretable systems* where users can mathematically analyze internal model mechanisms and *comprehensible systems* that emit symbols explaining user-driven explanations of how conclusions are reached. The study then culminates in the introduction of truly *explainable systems*, whereby automated reasoning is central to output crafted explanations without additional post-processing as a final step. This study expresses optimism that neural-symbolic learning would bridge the gap between connectionist and symbolic learning techniques and pave the way towards truly explainable AI systems.

Arrieta et al. 2020: This study offers an extensive review of existing XAI techniques for all kinds of machine-learning methods including deep learning. We consider this study as an extremely useful review that provides both machine-learning users and practitioners a better idea of existing XAI methodologies. Similar to the aforementioned study, this study expresses an interest in neural-symbolic paradigms for interpretable high-performance machine-learning systems.

1.1.1 Projected state of affairs

Given the current state of machine learning in NLP, it appears that deep learning models with millions of parameters dominate the field in terms of generalizable performance. Artificial Neural Networks (ANNs) are indeed powerful universal approximators and literature has shown that they can approximate arbitrary functions and logic programs. Approximation and generalization can be considered to be largely solved problems because of ANN learning and cross-validation.

$$\overbrace{\text{Approximation} \Rightarrow \text{Generalization}}^{\text{ANN learning + Cross-validation}} \Rightarrow \overbrace{\text{Semanticity and compositionality}}^{\text{Human interpretation}} \tag{1}$$

Recent research has shown the existence of multiple possible high-performance generalizable solutions to the ANN learning problem (Kepner et al., 2018). Given the existence of multiple solutions, the next frontier appears to be the process of selecting semantically relevant models; or otherwise models that actually reflect meaningful considerations which humans could easily pick out and which are not dependent on statistical artefacts.

1.2 Neural-backed decision trees

Wan et al. 2020: This is a powerful study which attempts to bridge the explainability of white-box decision trees with the universal approximation capabilities of neural networks. One limitation of this research appears to be the induced hierarchy in weight space. While this lends itself to high performance, this framework allows many of the "black-box" limitations of the ANN to persist; albeit in a hierarchical form. As a result, a probe into explainability would require careful analysis of the induced hierarchy and how adversarial samples would traverse such a hierarchy. Additionally, this framework would be difficult to implement for binary NLP tasks such as sentiment polarity detection where trees would be very short.

1.3 Inductive logic paradigms

Inductive learning of answer set programs proves to be a very useful machine learning technique since it is as explainable as can possibly be. Law et al. (2015) develop an inductive logic learning framework that can even handle noisy data sets. A limitation of this technique is the reliance on human-engineered search spaces along with atoms and variables. Conversely, a lack of human-driven feature engineering results in exponentially complex search spaces; thereby diminishing its utility as a general-purpose machine learning tool.

Evans and Grefenstette (2018) and Payani and Fekri (2019) both focus on reducing the reliance on human-driven feature engineering by integrating differentiable neural network components into their workflows. This has proven to be very useful and performant for relational classification tasks.

1.4 Neural-symbolic paradigms

Li et al. 2018: This study is an example of using a neural-symbolic paradigm to solve NLU entailment tasks. Here, the authors use neural networks to approximate symbolic functions such as the "find" function over sub-strings. Subsequently, they map together these neural modules through a logical framework which produces regular expressions as output rules.

General outlook: Based on literature review, it appears that much of research in this field requires the construction of features based on human input. While useful, this does not coincide strongly with the intent of this research; which is to achieve arbitrary explainable performance on NLP tasks without significant feature engineering.

1.5 Extraction of state-based automata from RNNs

Based on the previous sections, it appears that neural-backed decision trees are able to provide meaningful and quasi-explainable hierarchies for multi-label classifiers. Extending this into the textual domain with sequence classification tasks seems difficult and perhaps inappropriate, since usually output classifications are binary and final features tend to be sequential and token-dependent. An interesting analog to decision trees in the realm of time-series or sequences could be finite-state automata (FSA). Hou and Zhou (2018) implies the effectivity of extracting such FSAs directly from RNN models, which would be an interesting avenue for our research. From the outset, it appears that FSA and weighted finite automata (WFA) are both interpretable. Their state-based approach to modelling reflects some degree of explainability. The open question would be whether these models are robust to natural language.

1.6 Interpretable neural architectures

Instead of developing student-teacher algorithms which extract an interpretable surrogate model from an oracle, it could be seen as more efficient to work directly with an interpretable neural architecture. Schwartz et al. (2018) attempts to create a hybrid between RNN, CNN and weighed finite automata by analyzing surface-level string patterns. Peng et al. (2018) extend the aforementioned work by exploring more complex RNN architectures which are akint to WFSAs. Wang and Niepert (2019) focus on modifying common RNN architectures to include centroids which can be easily modeled as finite automata. Jiang et al. (2020) provide a variable RNN architecture which allows for initialization, modification and extraction using regular expressions.

References

A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

D. Doran, S. Schulz, and T. R. Besold. What does explainable ai really mean? a new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*, 2017.

R. Evans and E. Grefenstette. Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research*, 61:1–64, 2018.

B.-J. Hou and Z.-H. Zhou. Learning with interpretable structure from rnn. *arXiv preprint arXiv:1810.10708*, 2018.

C. Jiang, Y. Zhao, S. Chu, L. Shen, and K. Tu. Cold-start and interpretability: Turning regular expressions into trainable recurrent neural networks. 2020.

J. Kepner, V. Gadepally, H. Jananthan, L. Milechin, and S. Samsi. Sparse deep neural network exact solutions. In *2018 IEEE High Performance extreme Computing Conference (HPEC)*, pages 1–8. IEEE, 2018.

M. Law, A. Russo, and K. Broda. The ilasp system for learning answer set programs, 2015.

S. Li, H. Xu, and Z. Lu. Generalize symbolic knowledge with neural rule engine. *arXiv preprint arXiv:1808.10326*, 2018.

A. Payani and F. Fekri. Inductive logic programming via differentiable deep neural logic networks. *CoRR*, abs/1906.03523, 2019. URL `http://arxiv.org/abs/1906.03523`.

H. Peng, R. Schwartz, S. Thomson, and N. A. Smith. Rational recurrences. *arXiv preprint arXiv:1808.09357*, 2018.

R. Schwartz, S. Thomson, and N. A. Smith. Sopa: Bridging cnns, rnns, and weighted finite-state machines. *arXiv preprint arXiv:1805.06061*, 2018.

A. Wan, L. Dunlap, D. Ho, J. Yin, S. Lee, H. Jin, S. Petryk, S. A. Bargal, and J. E. Gonzalez. Nbdt: Neural-backed decision trees. *arXiv preprint arXiv:2004.00221*, 2020.

C. Wang and M. Niepert. State-regularized recurrent neural networks. *arXiv preprint arXiv:1901.08817*, 2019.