Explanation-Based Human Debugging of NLP Models: A Survey

Piyawat Lertvittayakumjorn and Francesca Toni

Department of Computing
Imperial College London, UK

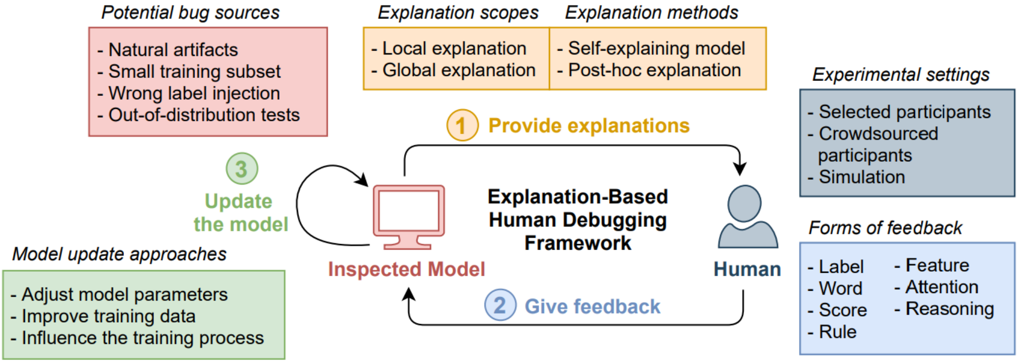
January 10, 2022

Overview

1. Introduction
2. Categorization of Existing Work
3. Human Factors
4. Open Problems

General Framework of EBHD

Explanaton-based human debugging (EBHD) : mitigating bugs using human feedback given in response to explanations



Contex

- tasks
  - text classification with single input (TC), natural language inference (NLI), question answering (QA)
- models
  - traditional models: naive Bayes (NB), logistic regression (LR), SVM
  - models involves word embeddings: CNN, fastText, Telling QA, Neural Operator (NeOp)
  - pre-trained language models: BERT, RoBERTa
- bug sources
  - natural artifacts (AR)
  - simulated: using a small subset for training (SS), injecting wrong labels (WL), out-of-distribution tests (OD), contaminating training data with decoys (in computer vision domain)

Workflow

providing explanations

- explanation scope
  - local explanations: demands a large amount of effort from feedback providers
  - global explanations: can hardly reveal details of complex model's inner workings
  - explanations for a group of predictions: no existing study for EBHD
- generating explanations
  - format: input-based explanations, example-based explanations, rule-based explanations, adversarial-based explanations ...
  - : how?: self-explaining (SE), post-hoc (PH) explanation methods
- presenting explanations
  - consider the background knowledge, desires and limits of the feedback providers
  - user-friendly, sound, complete, not overwhelming ...

Workflow

collecting feedback

- text classification:
  - decide which words are in fact relevant (WO), adjust word importance scores (WS), specifying relevancy scores for example-based explanations (ES), learned features (FE), learned rules (RU), check predicted labels or ground-truth labels
- table question answering:
  - identify where in the table and the question the model should focus (AT)
- complex tasks requiring reasoning:
  - compositional explanations to show how the humans would reason (RE) about the model's failure cases
- how to collect and utilize other forms of feedback

Workflow

updating the model

- adjust model parameters (M): important to make ensure that the adjustments made by humans generalize well to all examples
- improve training data (D): correcting mislabeled training examples, assigning noisy labels to unlabeled training examples, removing irrelevant words, creating augmented training examples
- influence the training process (T):
  - model specific: attention supervision, regularization, disabling learned features
  - model agnostic: user co-training,

No study testing which technique is more applicable for which task, domain, or model architecture

Workflow

iteration

- the debugging workflow can be done iteratively
- fix vital bugs first and finer bugs in later iterations
- avoid local decision pitfalls

Experimental setting

Setting	**advantages**	**disadvantages**
human participants (SP)	gain insights concerning human computer interaction	limited number of participants
crowdsourcing platform (SC)	conduct experiments at a large scale	quality control
simulation (SM)	faster and cheaper	may not reflect the effectiveness of the framework when deployed with real humans

Summary

Paper	Context			Workflow				Setting
	Task	Model	Bug sources	Exp. scope	Exp. method	Feed-back	Up-date	
Kulesza et al. (2009)	TC	NB	AR	G,L	SE	LB,WS	M,D	SP
Stumpf et al. (2009)	TC	NB	SS	L	SE	WO	T	SP
Kulesza et al. (2010)	TC	NB	SS	G,L	SE	WO,LB	M,D	SP
Kulesza et al. (2015)	TC	NB	AR	G,L	SE	WO,WS	M	SP
Ribeiro et al. (2016)	TC	SVM	AR	L	PH	WO	D	CS
Koh and Liang (2017)	TC	LR	WL	L	PH	LB	D	SM
Ribeiro et al. (2018b)	VQA TC	TellQA fastText	AR AR,OD	G	PH	RU	D	SP
Teso and Kersting (2019)	TC	LR	AR	L	PH	WO	D	SM
Cho et al. (2019)	TQA	NeOp	AR	L	SE	AT	T	NR
Khanna et al. (2019)	TC	LR	WL	L	PH	LB	D	SM
Lertvittayakumjorn et al. (2020)	TC	CNN	AR,SS,OD	G	PH	FE	T	CS
Smith-Renner et al. (2020)	TC	NB	AR,SS	L	SE	LB,WO	M,D	CS
Han and Ghosh (2020)	TC	LR	WL	L	PH	LB	D	SM
Yao et al. (2021)	TC	BERT*	AR,OD	L	PH	RE	D,T	SP
Zylberajch et al. (2021)	NLI	BERT	AR	L	PH	ES	D	SP

Model understanding

It is important to verify that the explanations help feedback providers form an accurate understanding of how the models work.

Examples

rule-based and keyword-based $>$ similarity-based, why $+$ why not $>$ why $>$ why not, interactive explanations $>$ static explanations

Examples

- some users did not understand explanations based on the absence of some words
- revealing inner workings could further help understanding but introduced additional workload

Willingness

what do humans naturally want to ?

- more than data labels, commonsense knowledge and English language knowledge
- neither complete nor precise, not quantitatively, selective, rarely refer to probabilities but express casual relationships

Trust

- increase trust:
  - showing more detailed explanations
  - the ability to provide feedback makes human trust
- decrease trust:
  - cannot increase human trust in high-stakes applications
  - explanations of low-quality models decrease trust
  - providing feedback decreases human trust

Frustration and Expectation

- frustration
  - increase frustration: ability to provide feedback
  - decrease frustration: poor explanations, too many details
- expectation
  - participants expected the model to improve after the interaction

Summary

- feedback providers: using developers or experts in the team be the providers; otherwise, collecting feedback implicitly
- explanations: avoiding forms which are difficult to understand; avoiding too much information
- feedback: relying on collective feedback; allowing to verify and modify
- update: showing the changes incrementally in real time

open problems

- beyond English text classification
- tackling more challenging bugs:
  - bugs happens less often
  - different people may give different feedback
  - injecting new knowledge with feedback
  - transfer techniques across modalities
- analyzing and enhancing efficiency: none of the selected studies considered the efficiency of three steps altogether, especially step3
- reliable comparison across paper: user studies are difficult to replicate
- towards deployment
  - integrating EBHD framework into available visualization systems
  - human-AI interaction guidelines and evaluate with potential end users

The End