Explanation-Based Human Debugging of NLP Models

Piyawat Lertvittayakumjorn

Imperial College London

pl1515@imperial.ac.uk

https://www.doc.ic.ac.uk/~pl1515/

Explainable AI (XAI)

- **Explainable AI** focuses on generating explanations for AI models as well as for their predictions.

- Given a model f and an input x,
  - **Local explanation** explains an individual prediction
    - Why does $f(x)$ equal a?
  - **Global explanation** explains the trained model independently of any specific prediction
    - How does f work?

Many Forms of Local Explanations in NLP

Task: Sentiment analysis
Input: Long and boring: I've read this book with much expectation, it was very boring all through out the book
Prediction: Negative

Form	Explanation
Rationale	boring
Saliency map	Long and boring: I've read this book with much expectation, it was very boring all through out the book
Counterfactual	Long and boring: I've read this book with much expectation, it was very awesome all through out the book
Training example	Stay Away: This just plain bad. Boring..... I did not find this the least bit entertaining nor interesting. It was a waste of my time. (**Label**: Negative)
Rule	If very ∧ boring, then **Prediction** = Negative (**Precision**: 94.7%)
Textual	The book did not meet the expectation, so the sentiment is negative.

Approaches for Global Explanations in NLP

- **Collection of Local Explanations**
  - Which local explanation method shall we use?
  - How shall we select a set of representative local explanations and how large is the set?
  - How shall we combine the local explanations to provide the global view of the model?

- **Surrogate Model**
  - An interpretable model which is trained to mimic the behavior of the complicated model.

If you want to know more,

- A Survey of the State of Explainable AI for Natural Language Processing (Danilevsky et al., 2020, AACL)
  - https://xainlp2020.github.io/xainlp/
- Interpreting Predictions of NLP Models (Wallace et al., 2020, EMNLP Tutorial)
  - https://www.youtube.com/watch?v=gprIzgIUW1s

Given the explanations, what's next?

XAI enables Human-AI Collaboration

- If an AI outperforms humans in a certain task, humans can learn and distill knowledge from the given explanations.
- If an AI's performance is close to human intelligence, the explanations can increase humans' trust in the AI and enable human-AI negotiations.
- If an AI is duller than humans, the explanations help humans verify the decisions made by the AI and also improve the AI.

Explanation-Based Human Debugging (EBHD) of NLP Models

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Muller. 2018. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. ITU Journal

Outline

- General EBHD framework
- A few instances of the framework
- Specific components of the framework
- Research on Human Factors
- Open problems
- Conclusion

Main reference of this talk

Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-Based Human Debugging of NLP Models: A Survey. TACL (Forthcoming).

Bug & Debugging

- What is a bug in machine learning?
  1. An implementation error, similar to a software bug (Selsam et al., 2017)
  2. A particularly damaging or inexplicable test error (Cadamuro et al., 2016)
  3. Contamination in the learning and/or prediction pipeline that makes the model produce incorrect predictions or learn error-causing associations (Adebayo et al., 2020)

- In this talk, we adopt the last definition (so called a model bug).
  - E.g., spurious correlation, labelling errors, and undesirable behavior in out-of-distribution (OOD) testing.

Bug & Debugging

- What is debugging in Machine Learning?
  1. A process of identifying or uncovering bugs which are causes of model errors
  2. Identifying the bugs + fixing or mitigating them

- In this talk, we adopt the second definition.
- **Explanation-Based Human Debugging (EBHD)** = The process of fixing or mitigating bugs in a trained model using human feedback given in response to explanations for the model.

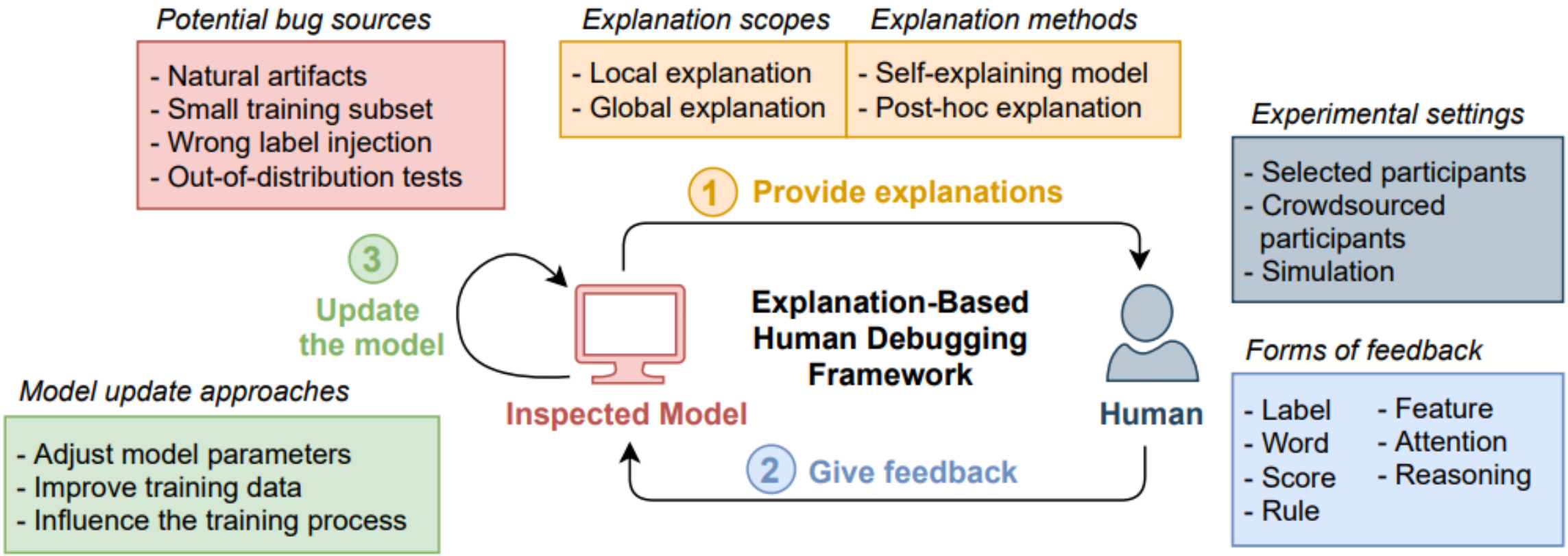
Outline

- General EBHD framework
- A few instances of the framework
- Specific components of the framework
- Research on Human Factors
- Open problems
- Conclusion

Main reference of this talk

Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-Based Human Debugging of NLP Models: A Survey. TACL (Forthcoming).

General EBHD framework



Outline

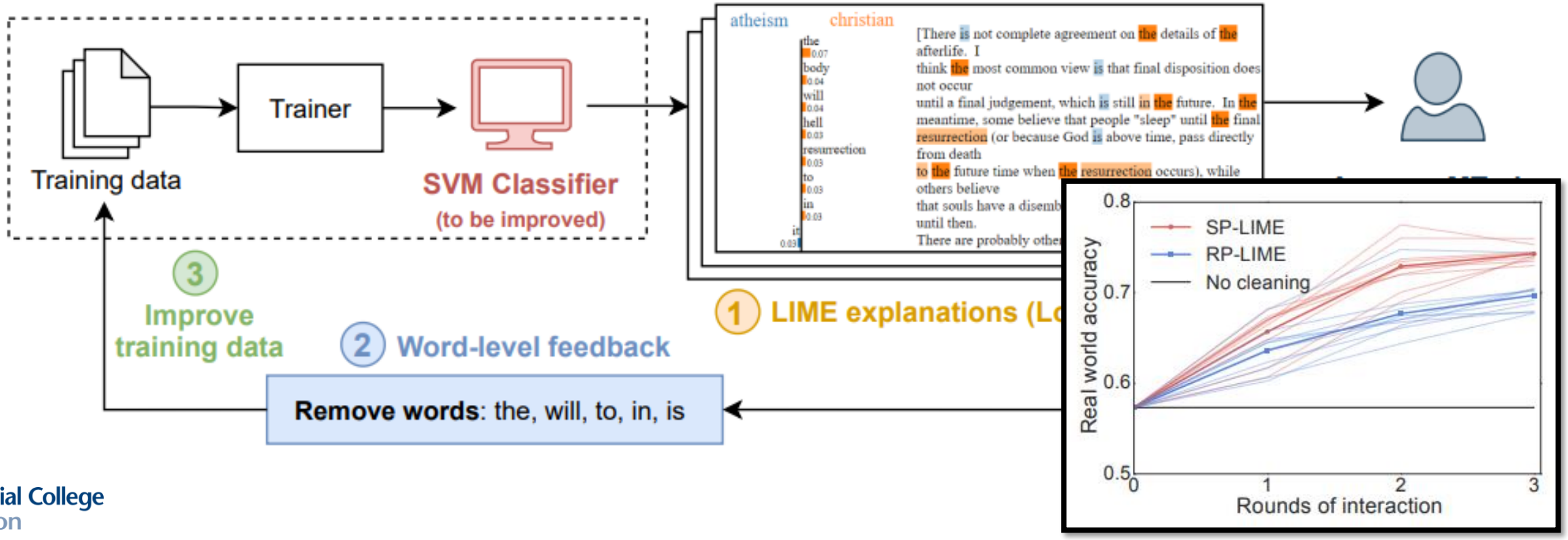
- General EBHD framework
- A few instances of the framework
- Specific components of the framework
- Research on Human Factors
- Open problems
- Conclusion

Main reference of this talk

Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-Based Human Debugging of NLP Models: A Survey. TACL (Forthcoming).

Example 1: LIME (Ribeiro et al., 2016)

- LIME = Local Interpretable Model-agnostic Explanations
- **Context**: Text classification, 20Newsgroups (Atheism vs Christian)



Example 2: EluciDebug (Kulesza et al., 2015)

- **Context:**
  - Text classification
  - 20Newsgroups (Hockey vs Baseball)
  - **Model:** Multinomial Naïve Bayes

$$\hat{y} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{l=1}^{L} P(x_{il}|c)$$

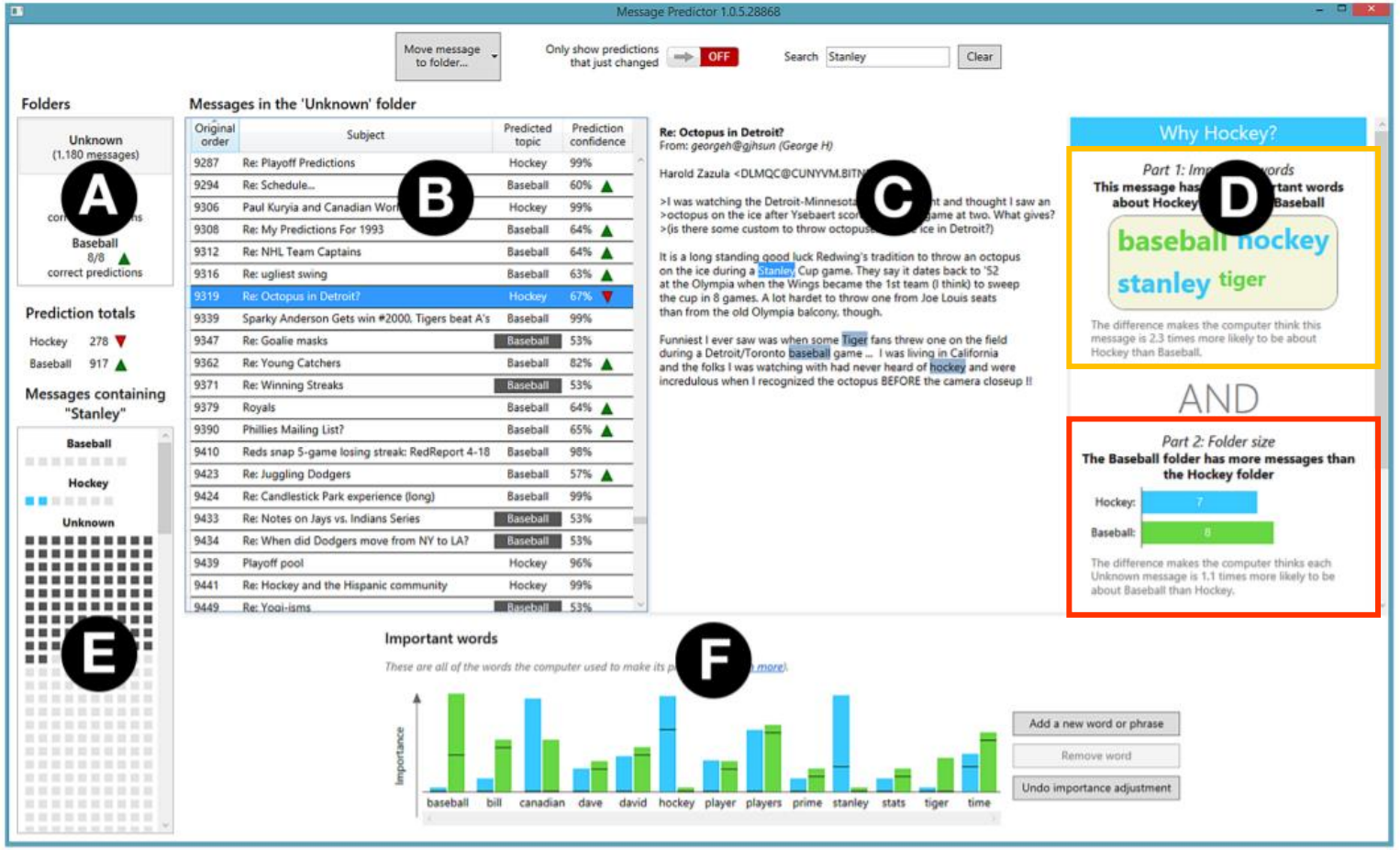
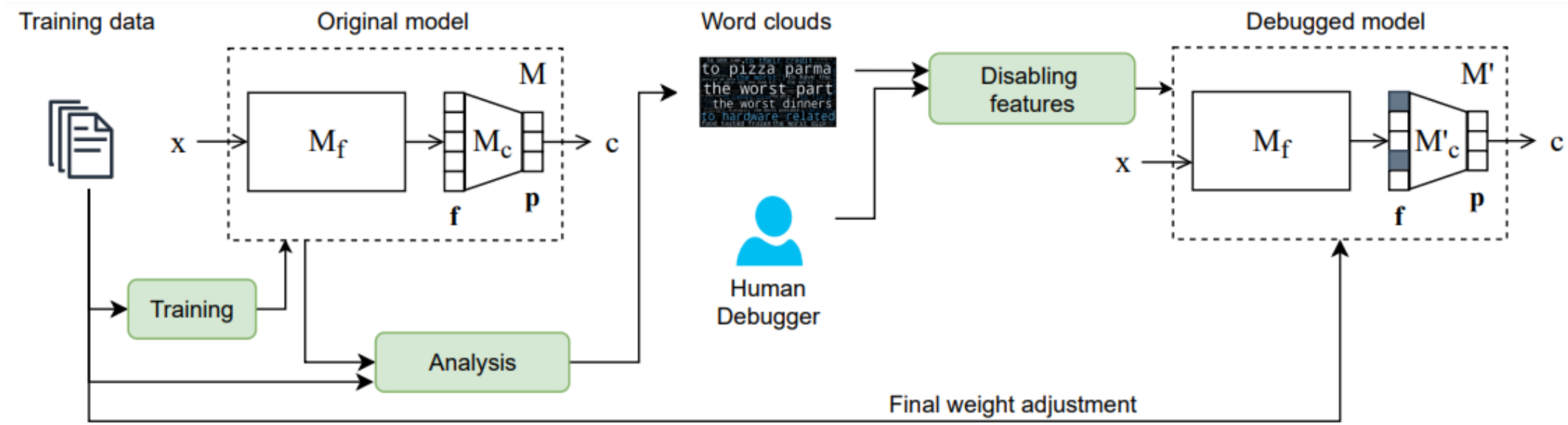


Figure 1. The EluciDebug prototype. (A) List of folders. (B) List of messages in the selected folder. (C) The selected message. (D) Explanation of the selected message's predicted folder. (E) Overview of which messages contain the selected word. (F) Complete list of words the learning system uses to make predictions.

Example 3: FIND (Lertvittayakumjorn et al., 2020)

- **FIND** = Feature Investigation and Disabling
- **Context**: Text classification, Six datasets, **Model**: 1D CNNs



Outline

- General EBHD framework
- A few instances of the framework
- Specific components of the framework
- Research on Human Factors
- Open problems
- Conclusion

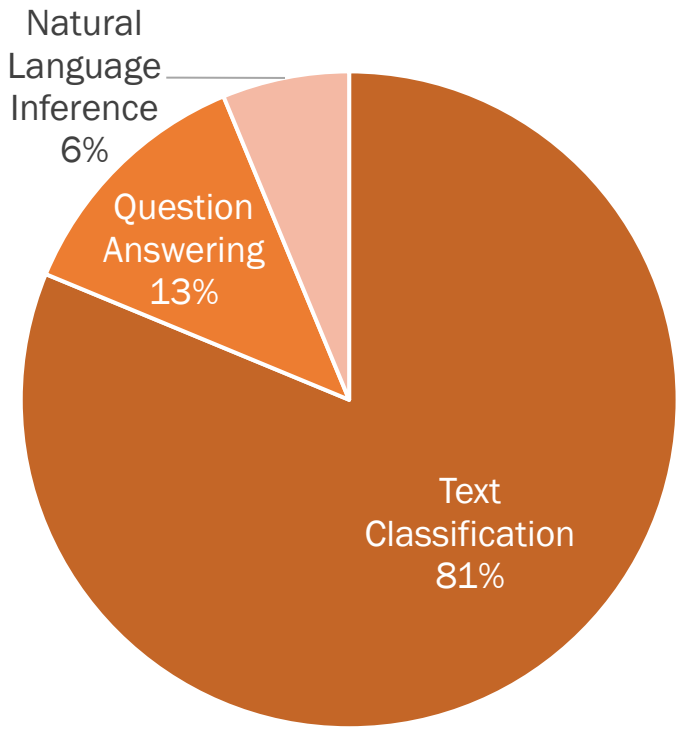
Main reference of this talk

Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-Based Human Debugging of NLP Models: A Survey. TACL (Forthcoming).

Bug Context – Task

Most researchers work on text classification because, for this task, it is much easier for lay participants to understand explanations and give feedback (Ghai et al., 2021).

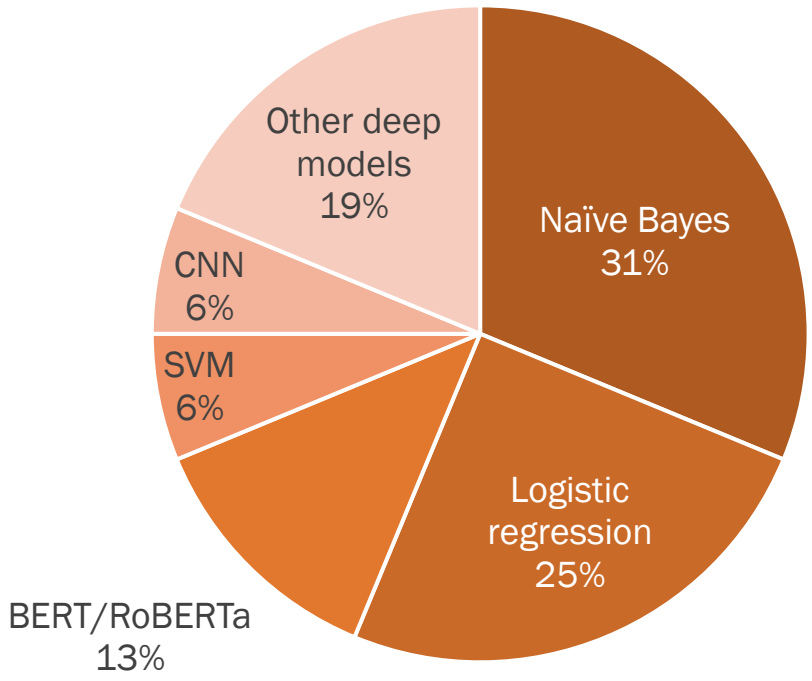
Target NLP task



- Text classification
  - Topic classification
  - Spam classification
  - Sentiment analysis
  - Auto-coding of transcripts
  - …
- Question answering
  - Visual question answering
  - Table question answering
- Natural language inference

Bug Context – Model

Inspected model



- Interpretable models
  - Naïve Bayes
  - Logistic Regression
- Black-box models
  - SVM with RBF kernel
  - fastText (Joulin et al., 2017)
  - 1D CNN (Kim, 2014)
  - TellingQA (Zhu et al., 2016)
  - NeOp (Cho et al., 2018)
  - BERT (Devlin et al., 2019)
  - RoBERTa (Liu et al., 2019)

Bug Context – Bug Sources

- Natural bugs – Natural artifacts
  - The input texts have signals which are correlated to but not the reasons for specific outputs.

Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	**The doctor** was **paid** by **the actor**. $\xrightarrow[\text{WRONG}]{}$ The doctor paid the actor.
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near **the actor danced**. $\xrightarrow[\text{WRONG}]{}$ The actor danced.
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If **the artist slept**, the actor ran. $\xrightarrow[\text{WRONG}]{}$ The artist slept.

Table 1: The heuristics targeted by the HANS dataset, along with examples of incorrect entailment predictions that these heuristics would lead to.

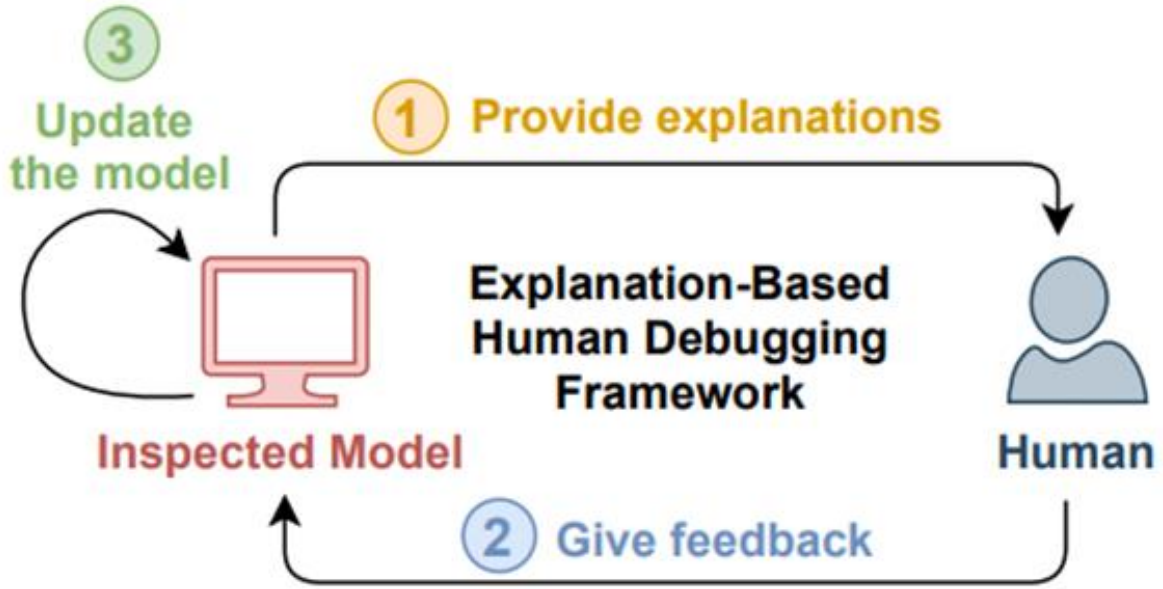
Thomas McCoy, Ellie Pavlick, Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. ACL

Bug Context – Bug Sources

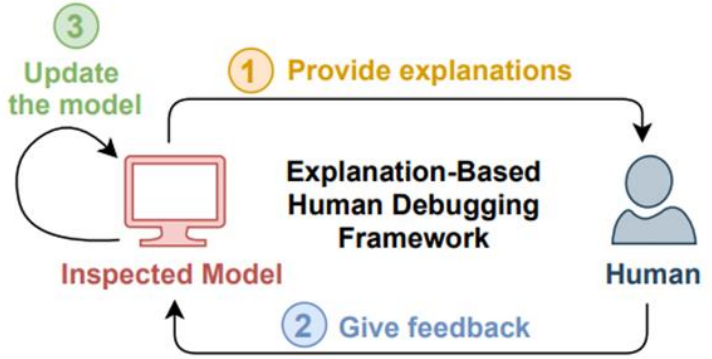
- Simulated bugs
  - Contaminating input texts in the training data with decoys (i.e., injected artifacts)
  - Using a small subset of labeled data for training
  - Injecting wrong labels into the training data
  - Using out-of-distribution tests

The Workflow

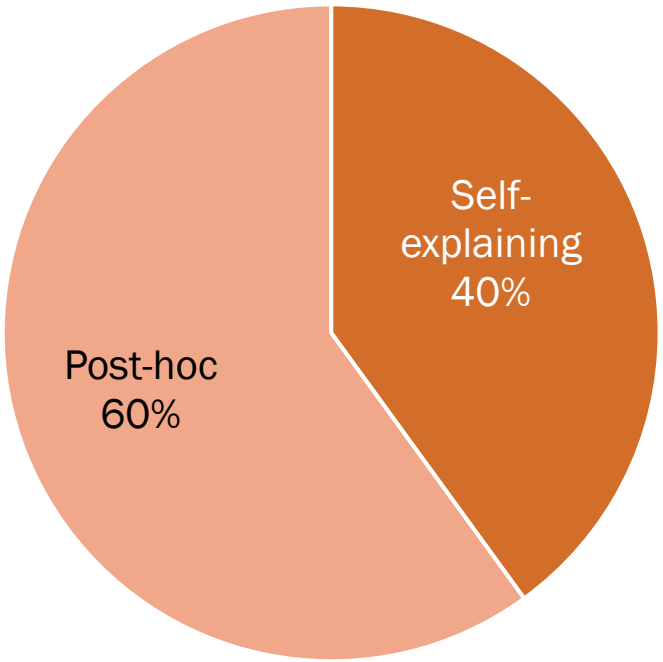
- The three steps need to be decided harmoniously to create an effective debugging workflow.



Step 1: Provide Explanations

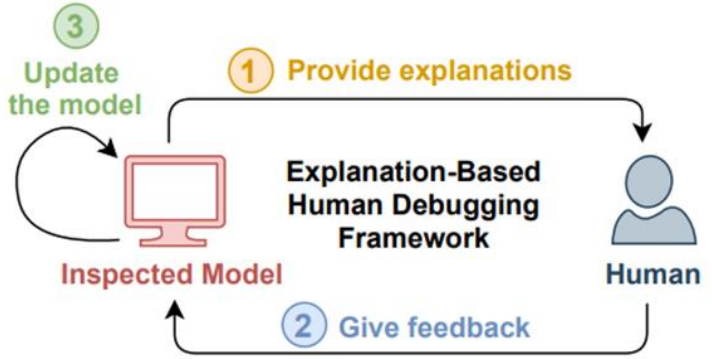


Explainability approach / form

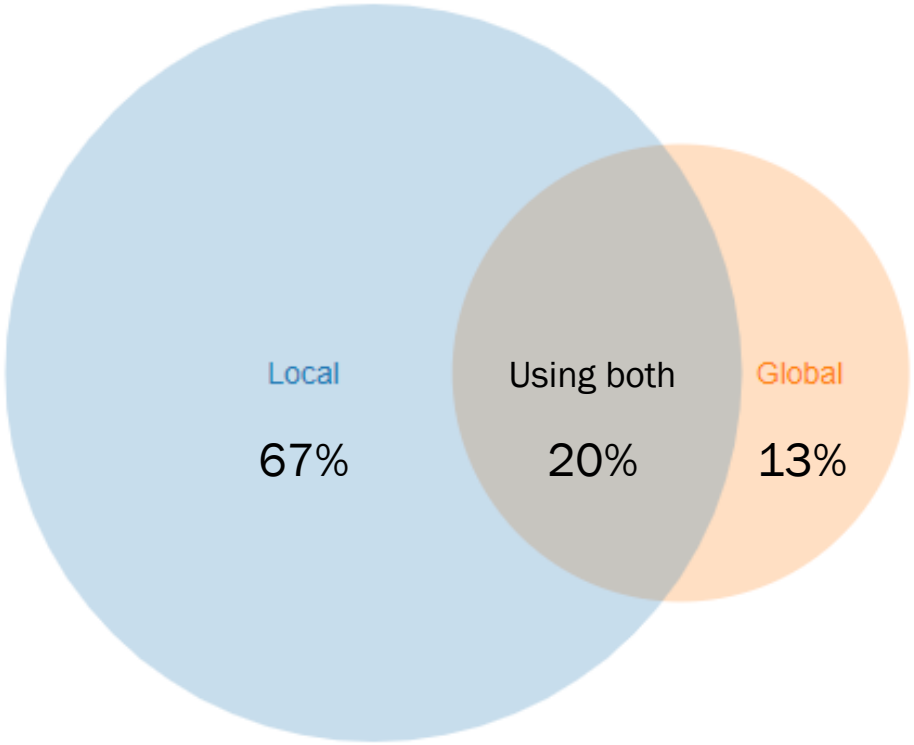


- **Self-explaining approach**
  - Local explanations can be obtained at the same time as predictions.
  - Global explanations are often the models themselves.
- **Post-hoc approach**
  - Performs additional steps to extract explanations.
  - Input-based >> Example-based > Others

Step 1: Provide Explanations



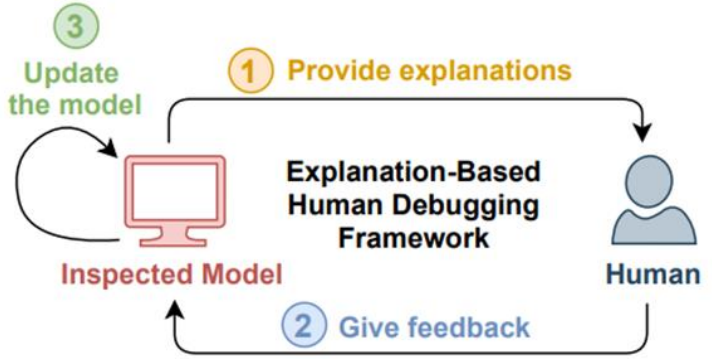
Explainability scope



- **Global explanations**
  - Reveal significant bugs
- **Local explanations**
  - Reveal fine-grained bugs
  - Need a strategy to pick examples to explain
    - Incorrect predictions
    - Non-redundancy
    - Informativeness criteria

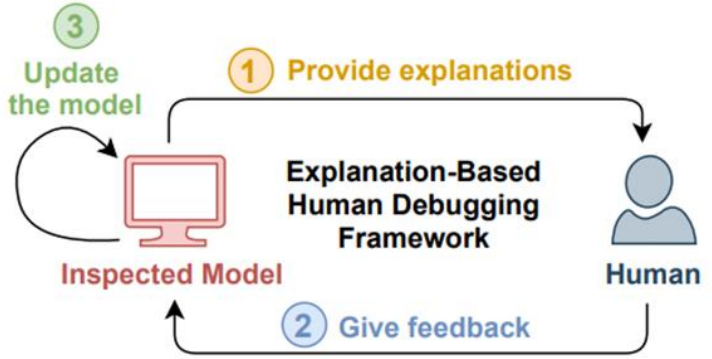
Chart generated by https://www.meta-chart.com/venn

Step 2: Collect Feedback

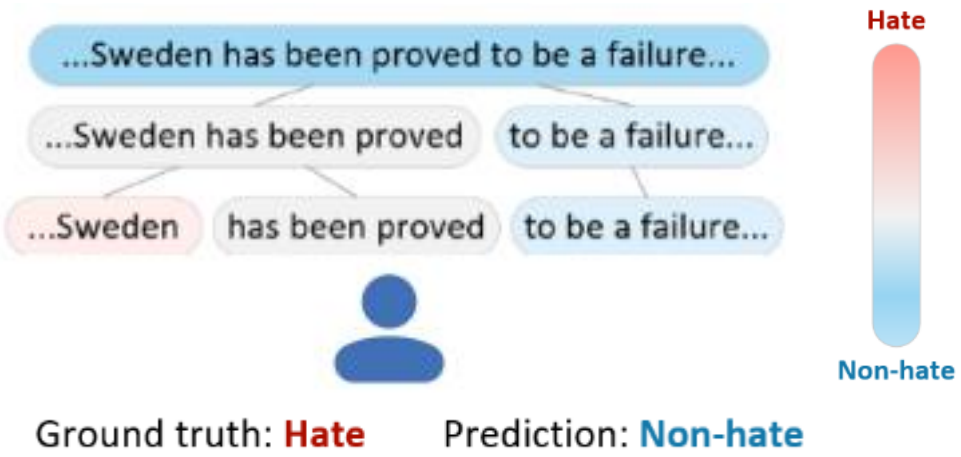


Explanation Form	Feedback Method
Input-based explanations • Rationales • Relevance scores, Saliency maps • Hierarchical heat maps	• Identify whether words/tokens/cells are relevant or not • Adjust the word importance scores • Explain proper reasoning via rules
Example-based explanations • Influential training examples	• Provide correct labels • Provide relevancy scores
Global explanations • Learned features • Adversarial rules	• Check whether the learned features are relevant • Check whether the adversarial rules are semantically equivalent

Step 2: Collect Feedback



- **Task**: Hate speech detection (Yao et al., 2021)
- **Explanation** (Hierarchical heat-map)



Interaction describes how the importance of a phrase changes when the other word or phrase is absent or present.

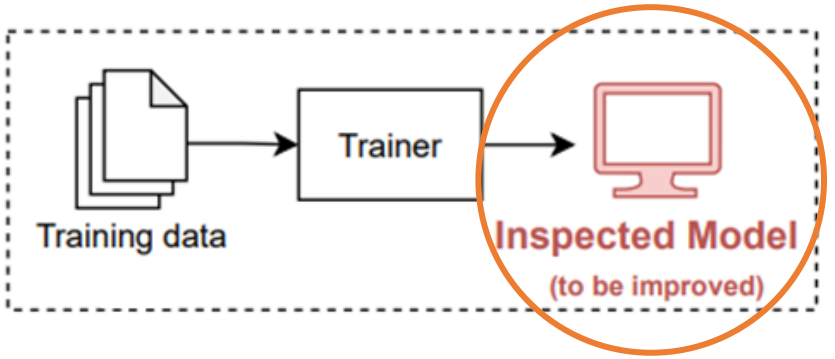
Feedback (Reasoning rule)

Because the word *"Sweden"* is a country, *"failure"* is negative, and *"Sweden"* is less than 3 dependency steps from *"failure"*, Attribution score of *"Sweden"* should be decreased. Attribution score of *"failure"* should be increased. The interaction score of *"Sweden"* and *"failure"* should be increased.

@Is(Word1, country)
∧ @Is(Word2, negative)
∧ @LessThan(Word1, Word2) →
DecreaseAttribution(Word1)
∧ IncreaseAttribution(Word2)
∧ IncreaseInteraction(Word1, Word2).

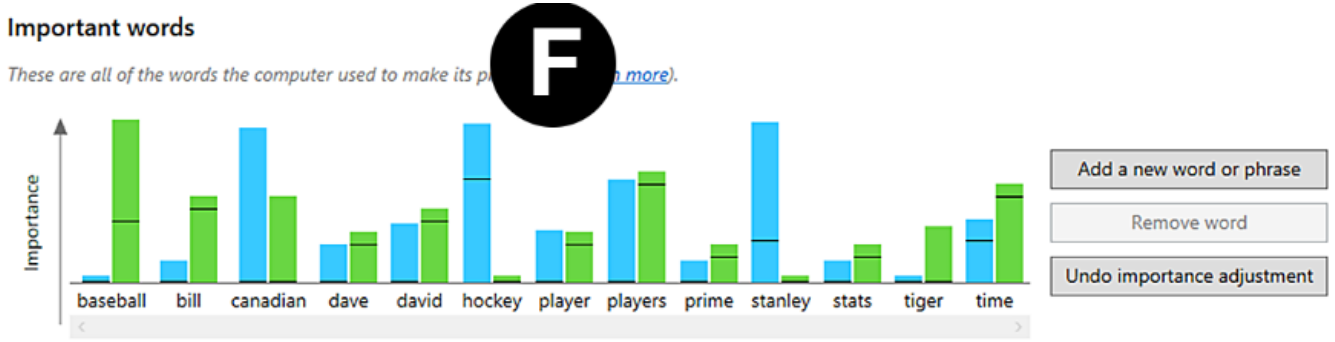
Huihan Yao, Ying Chen, Qinyuan Ye, Xisen Jin, Xiang Ren. 2021. Refining Language Models with Compositional Explanations. arXiv: 2103.10415 ; https://shanzhenren.github.io/talk/Ren-talk%202021-Mar-PDF.pdf

Step 3: Update the Model



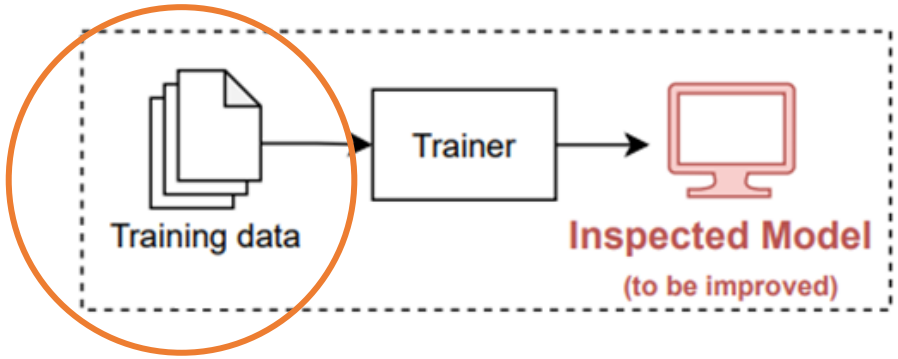
(1) Directly adjust the model parameters

- Suitable for transparent models where the explanation displays the model parameters in an intelligible way
- Fast as it does not require retraining the model
- How can we ensure that the adjustments made by humans generalize well to all examples?
  - Showing the changes to users in real-time (metrics, predictions, explanations)
  - Allowing the users to undo their actions where the results are not desirable



Todd Kulesza, Margaret Burnett, Weng-Keen Wong, Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. IUI.

Step 3: Update the Model

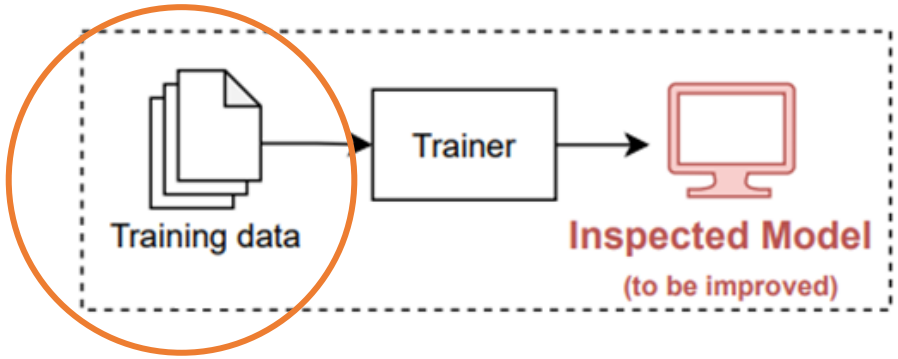


(2) Improving the training data

- Removing irrelevant words from input texts (Ribeiro et al., 2016)
- Correcting mislabeled training examples (e.g., Koh and Liang, 2017)
- Adding more training examples by
  - Assigning (noisy) labels to unlabeled examples (Yao et al., 2021)
  - Creating augmented training examples to reduce the effects of the artifacts (e.g., Teso and Kersting, 2019; Zylberajch et al., 2021)

- This approach works at the training data, so it is model-agnostic.

Step 3: Update the Model



CAIPI (Teso and Kersting, 2019)

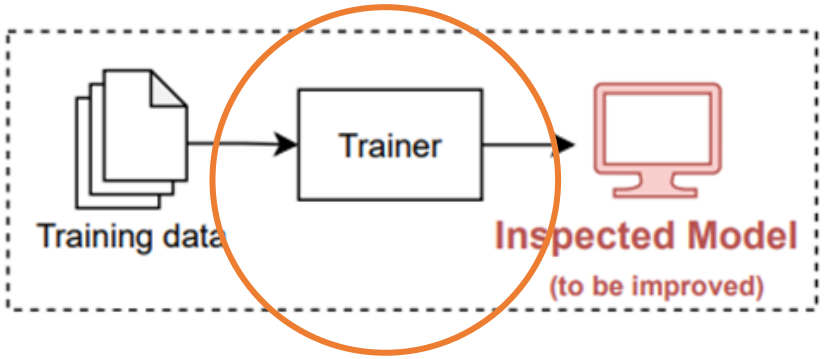
- Use active learning to pick a query
- Predict and explain the query
- Collect feedback from the users
  - True label
  - Irrelevant features in the explanation
- Generate counterexamples
  - Vary the removed feature but retain the label
- Combine the augmented data with the original dataset and retrain the model

Algorithm 1 CAIPI takes as input a set of labelled examples $\mathcal{L}$, a set of unlabelled instances $\mathcal{U}$, and iteration budget T.

$f \leftarrow \text{FIT}(\mathcal{L})$
repeat
  $x \leftarrow \text{SELECTQUERY}(f, \mathcal{U})$
  $\hat{y} \leftarrow f(x)$
  $\hat{z} \leftarrow \text{EXPLAIN}(f, x, \hat{y})$
  Present x, $\hat{y}$, and $\hat{z}$ to the user
  Obtain y and explanation correction $\mathcal{C}$
  $\{(\bar{x}_i, \bar{y}_i)\}_{i=1}^{c} \leftarrow \text{TOCOUNTEREXAMPLES}(\mathcal{C})$
  $\mathcal{L} \leftarrow \mathcal{L} \cup \{(x, y)\} \cup \{(\bar{x}_i, \bar{y}_i)\}_{i=1}^{c}$
  $\mathcal{U} \leftarrow \mathcal{U} \setminus (\{x\} \cup \{\bar{x}_i\}_{i=1}^{c})$
  $f \leftarrow \text{FIT}(\mathcal{L})$
until budget T is exhausted or f is good enough
return f

Stefano Teso & Kristian Kersting. Explanatory interactive machine learning. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 239-245).

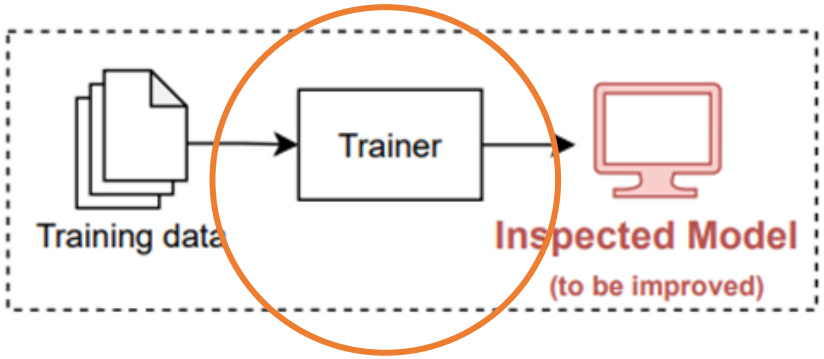
Step 3: Update the Model



(3) Influence the (re)training process

- Aim to make the resulting model behave as the feedback suggests
  - Feature disabling (Lertvittayakumjorn et al., 2020)
  - Regularizing the explanations, e.g., attention scores (Cho et al., 2018), integrated gradients (Yao et al., 2021)
  - Constraint optimization (Stumpf et al., 2009)
  - User co-training (Stumpf et al., 2009)

Step 3: Update the Model



Explanation Regularization (Yao et al., 2021)

- Regularize attribution scores

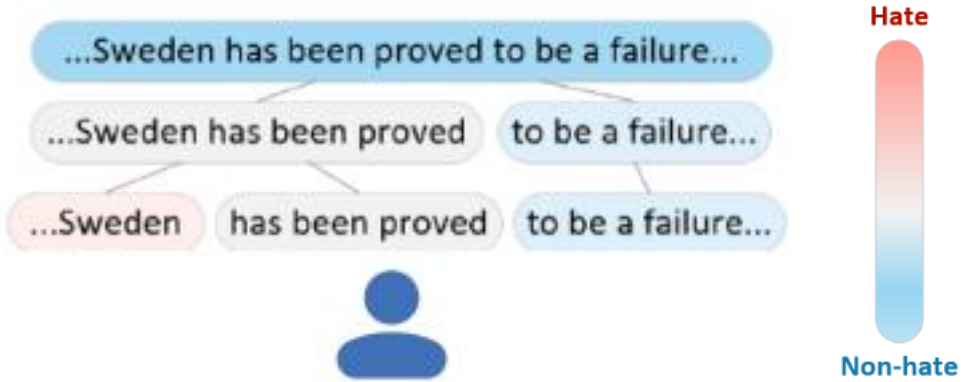
$$\mathcal{L}^{attr} = \sum_{c}^{C} \sum_{p \in \mathcal{R}} (\phi^c(p; \boldsymbol{x}) - t_p^c)^2$$

- Regularize interaction scores

$$\mathcal{L}^{inter} = \sum_{c}^{C} \sum_{\{p,q\} \in \mathcal{R}} (\varphi^c(p, q; \boldsymbol{x}) - \tau_{p,q}^c)^2$$

- Total loss

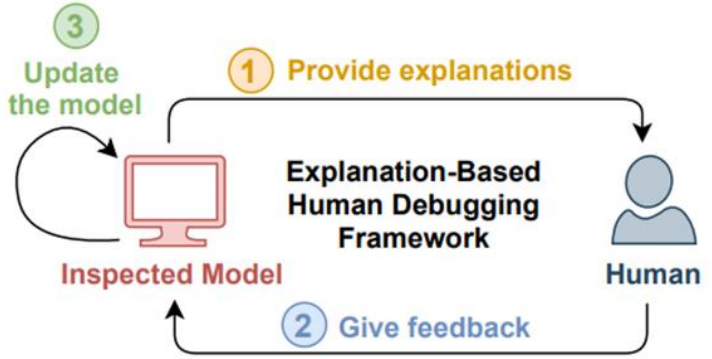
$$\mathcal{L} = \mathcal{L}' + \alpha(\mathcal{L}^{attr} + \mathcal{L}^{inter})$$



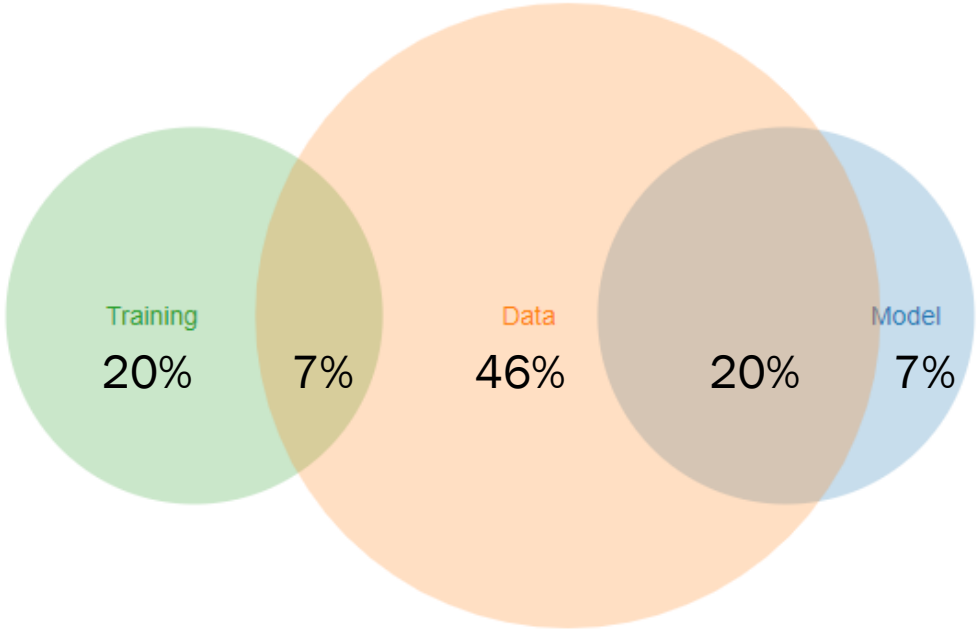
Attribution score of "**Sweden**" should be *decreased*.
Attribution score of "**failure**" should be *increased*.
The interaction score of "**Sweden**" and "**failure**" should be *increased*.

Huihan Yao, Ying Chen, Qinyuan Ye, Xisen Jin, Xiang Ren. 2021. Refining Language Models with Compositional Explanations. arXiv: 2103.10415 ; https://shanzhenren.github.io/talk/Ren-talk%202021-Mar-PDF.pdf

Step 3: Update the Model



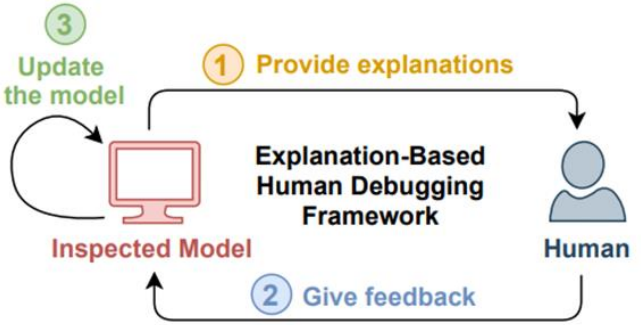
Update Approach



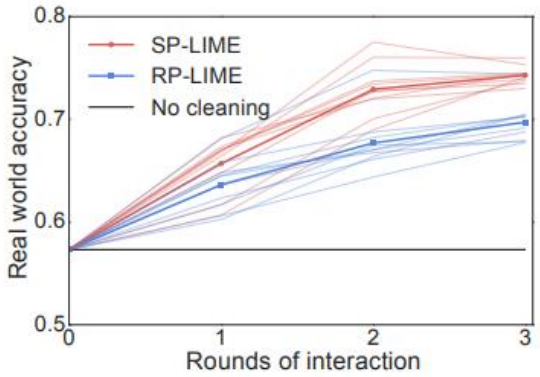
- Most of the existing works update the model by improving the training data, some of which combine it with one of the other two approaches.

Chart generated by https://www.meta-chart.com/venn

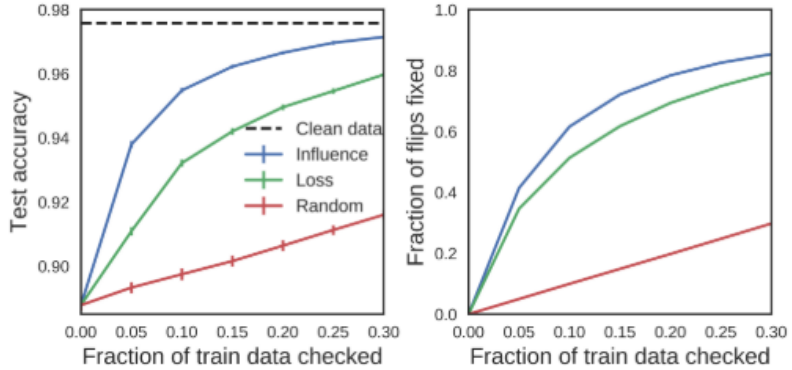
Iterative improvement



- **Iterative improvement** can be done if the explanations change after model update.
  - This allows humans to fix vital bugs first and finer bugs in later iterations.
    - (Ribeiro et al., 2016)

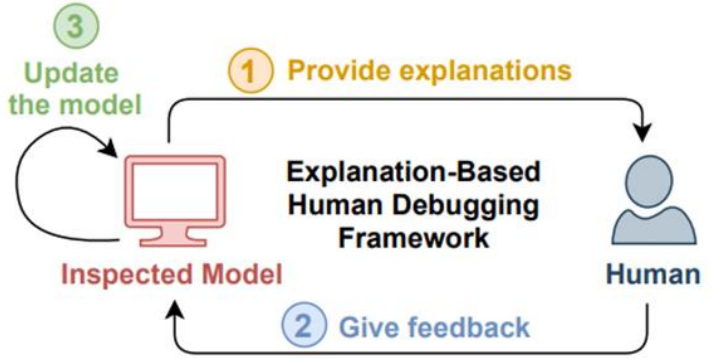


(Koh and Liang, 2017)

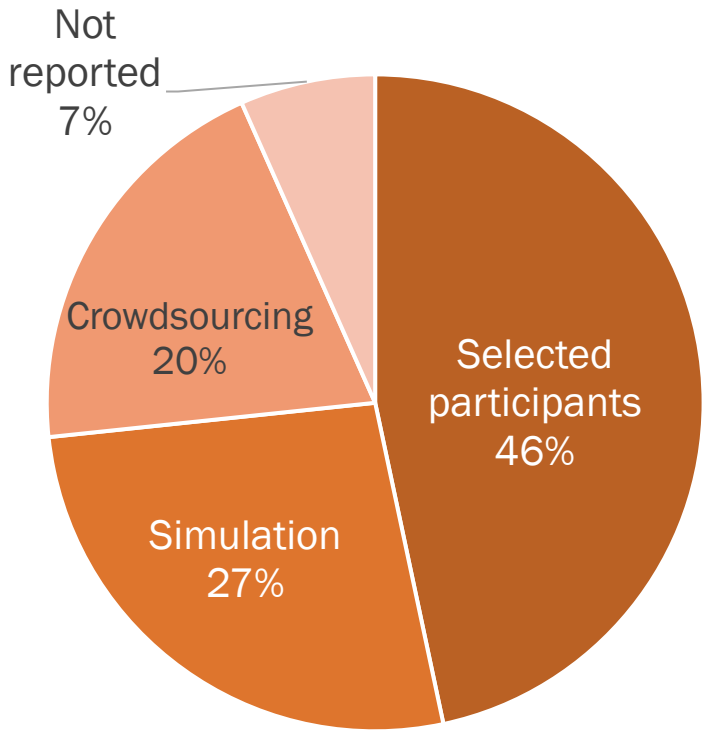


  - It may be susceptible to *local decision pitfalls* where local improvements could add up to inferior overall performance (Wu et al., 2019)

Experimental Setting



Feedback providers



- **Selected participants**
  - E.g., domain experts, NLP experts, students
  - Could be done in-person
- **Crowdsourcing**
  - Mostly via Amazon Mechanical Turk
  - Need some quality control
- **Simulation (No humans involved)**
  - Use oracles as human feedback
  - Faster and cheaper but not natural

Outline

- General EBHD framework
- A few instances of the framework
- Specific components of the framework
- Research on Human Factors
- Open problems
- Conclusion

Main reference of this talk

Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-Based Human Debugging of NLP Models: A Survey. TACL (Forthcoming).

Model Understanding

- Some forms of explanations are more effective for creating good user understanding of the models than others.
  - Rules & keywords (are better than similar examples) (Stumpf et al., 2009)
  - Explaining why (is better than explaining why not) (Lim et al., 2009)
  - Interactive explanations (are better than static ones) (Cheng et al., 2019)

- Our suggestions
  - Avoid using forms of explanations which are difficult to understand
  - Allow the feedback providers to request more information if they are interested

Human Feedback Characteristics

- Human feedback is not always complete, correct, or accurate (Ghai et al., 2021).
  - It focuses on a few features that are most different from human expectation, ignoring the others.
  - It is usually not accurate for correcting model explanations quantitatively.

- **Our suggestions**
  - Rely on collective feedback rather than individual feedback
  - Allow feedback providers to verify and modify their feedback before applying it to update the model

Trust & Frustration

- Explanations of low-quality models decrease trust and cause frustration to the users. Also, it is inconclusive whether an ability to provide feedback makes human trust and acceptance rally (Smith-Renner et al., 2020; Honeycutt et al., 2020).

- **Our suggestions**
  - If possible, let the developers or domain experts in the team (rather than end users) be the feedback providers.
  - Collect end users' feedback implicitly or collect without telling them that the explanations are of the production system.

Expectation

- Some humans expected the model to improve after the session where they interacted with the model, regardless of whether they saw explanations or gave feedback during the interaction session (Smith-Renner et al., 2020).

- **Our suggestions**
  - Display the improvements after the model gets updated.
  - Where possible, show the changes incrementally in real time, allowing the feedback providers to check if their feedback works as expected or not.

Outline

- General EBHD framework
- A few instances of the framework
- Specific components of the framework
- Research on Human Factors
- Open problems
- Conclusion

Main reference of this talk

Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-Based Human Debugging of NLP Models: A Survey. TACL (Forthcoming).

Open Problems

- Beyond English Text Classification
- Tackling More Challenging Bugs
  - Dealing with conflicting pieces of feedback
  - Injecting new knowledge to the model
- Analyzing and Enhancing Efficiency
- Reliable Comparison across Papers
- Towards Deployment

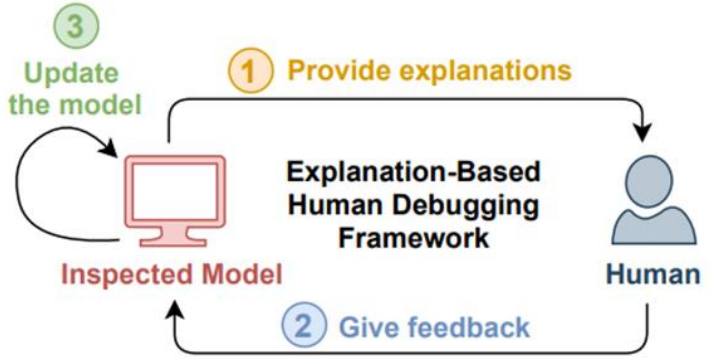
Outline

- General EBHD framework
- A few instances of the framework
- Specific components of the framework
- Research on Human Factors
- Open problems
- Conclusion

Main reference of this talk

Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-Based Human Debugging of NLP Models: A Survey. TACL (Forthcoming).

Conclusion



- EBHD process makes explanations actionable.
- 3 main steps: Explain → Feedback → Update (Repeat)
  - They should be designed harmoniously with respect to the bug context, i.e., the task, the model, and the bug sources.
- Humans are not perfect oracles. Please take care of them.
- Many challenges are still not fully solved, e.g., generalizability, efficiency, comparison, and deployment.

Thank you / Q&A

Piyawat Lertvittayakumjorn
Imperial College London
pl1515@imperial.ac.uk | @plkumjorn
https://www.doc.ic.ac.uk/~pl1515/