Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing

Sarah Wiegreffe and Ana Marasovic

School of Interactive Computing
Georgia Institute of Technology

January 11, 2022

Overview

1. Introduction
2. Link Between EXNLP Data, Modeling, and Evaluation Assumptions
3. Rise of Structured Explanations
4. Increasing Explanation Quality

Goals of human justification

1. to aid models with additional training supervision
2. to train interpretable models that explain their own predictions
3. to evaluate plausibility of model-generated explanations

Three types of explanations

Instance	Explanation
Premise: A white race dog wearing the number eight runs on the track. *Hypothesis:* A white race dog runs around his yard. *Label:* contradiction	`(highlight)` *Premise:* A white race dog wearing the number eight runs on the track . *Hypothesis:* A white race dog runs around his yard . `(free-text)` A race track is not usually in someone's yard.
Question: Who sang the theme song from Russia With Love? *Paragraph:* ...The theme song was composed by Lionel Bart of Oliver! fame and sung by Matt Monro... *Answer:* Matt Monro	`(structured)` *Sentence selection:* (not shown) *Referential equality:* "the theme song from russia with love" (from question) = "The theme song" (from paragraph) *Entailment:* X was composed by Lionel Bart of Oliver! fame and sung by ANSWER. ⊢ ANSWER sung X

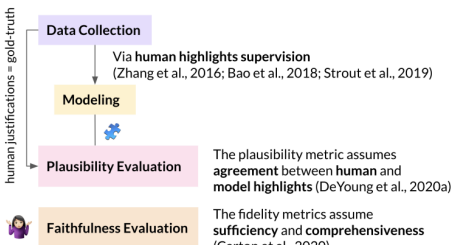
Properties of highlights

Instance with Highlight	Highlight Type Clarification
Review: this film is extraordinarily horrendous and I'm not going to waste any more words on it. *Label*: negative	(¬comprehensive) *Review:* this film is [redacted] and I'm not going to waste any more words on it.
Review: this film is extraordinarily horrendous and I'm not going to waste any more words on it . *Label*: negative	(comprehensive) *Review:* this film is [redacted] and I'm not going to [redacted].
Premise: A shirtless man wearing white shorts. *Hypothesis:* A man in white shorts is running on the sidewalk. *Label*: neutral	(¬sufficient) *Premise:* [redacted] *Hypothesis:* [redacted] man [redacted] running on the sidewalk.

structure explanations

- chains of facts: detail the reasoning steps to reach an answer
- semi-structured text: place constraints on the textual explanations that annotators can write
- explanation graphs: combination of chains of facts and semi-structured text

Supervised models' development



definition

- plausibility: according to humans, how well a highlight supports a predicted label
- faithfulness or fidelity: how accurately a highlight represents the model's decision process

human-annotated highlights are used only for evaluation of plausibility but not faithfulness

Supervised models' development

Sufficiency is necessary.

Examples

Neutral E-SNLI: not justifiable by highlight

Premise: A shirtless man wearing white shorts. *Hypothesis:* A man in white shorts is running on the sidewalk. *Label*: neutral

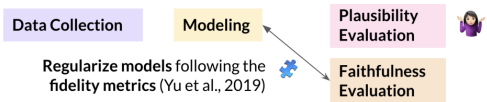
(¬sufficient) *Premise:* [redacted] *Hypothesis:* [redacted] man [redacted] running on the sidewalk.

Examples

No-attack WIKIATTACK: the absence cannot be highlighted

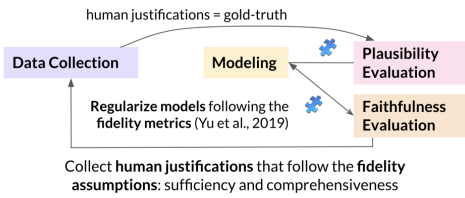
1. avoiding human-annotated highlights with low sufficiency
2. assessing whether the true label can be explained by highlighting

Unsupervised models' development



measurement and modeling of faithfulness cannot influence how human-authored explanantions should be collected

Recommended unsupervised model's development



Requires comprehensiveness.

- post-hoc assessment of comprehensiveness from a general description of data collection is error-prone $\rightarrow$ precisely report how explanations were collected
- not popularize data collection decisions as universally necessary $\rightarrow$ better documentation
- non-comprehensiveness can hinder evaluating plausibility of comprehensive model highlights

structured explanations

template-like free-text explanations

Examples

- "There is <hypothesis>"
- "<answer> is the only option that is correct/obvious"

uninformative, can result in artifact-like behaviours

- uninformative, can result in artifact-like behaviours

Are all template-like explanations uninformative? running pilot studies to explore how people define and generate explanations for a task

- informative human explanations are naturally structured
  - embracing the structure, consulting domain experts or follow literature
  - highlight in a dataset datasheet
- do not reveal any obvious structure
  - do best to control the quality

quality control

- two-stage approaches: collect-and-judge or collect-and-edit
- teach and test the underlying task
- addressing ambiguity: collect both labels and explanations from the same annotators or include a checker

Increasing explanation diversity

- use a large set of annotators
- multiple annotations per instance
- add contrastive and negative explanations
  - no dataset that contains contrastive free-text or strutured explanations
    - explanations answering the question "why ... instead of ..."
    - explanations for other labels besides the gold label
  - explanations that are invalid for an (input, gold label) pair
    - low-scoring instances or instances pre-editing

The End