# Towards Conversational Data Annotation: Personalized Annotation Explanation Generation via Large Language Models

## Literature Review & Challenges

Mark Nagengast Porro

# Contents

# Are Human Explanations Always Helpful? Towards Objective Evaluation of Human Natural Language Explanations I/V

## Summary

- Human-annotated explanations commonly used as ground truth for metrics such as *BLEU* and *ROUGE*
  - Semantic similarity or word-matching generally does not suffice to judge an explanation's validity
  - Often subjective or task-dependent
- Metric for evaluating quality of free-text human-annotated explanation for labeling of an instance, called *TREU*
- The quality of an explanation is derived from its helpfulness to a model's prediction performance
- Helpfulness judged both during fine-tuning and inference, expanding on *Simulatability* score

# Are Human Explanations Always Helpful? Towards Objective Evaluation of Human Natural Language Explanations II/V

## Summary

- *Simulatability* measures the difference in performance when including and excluding the explanation in the input

| | |
|---|---|
| Baseline | **explain:** *Question Content* **choice1:** *Choice1* **choice2:** *Choice2* **choice3:** *Choice3* |
| Infusion | **explain:** *Question Content* **choice1:** *Choice1* **choice2:** *Choice2* **choice3:** *Choice3* **<sep> because** *Explanation* |

Figure: Two settings: No explanations (Baseline) and explanations as part of input (Infusion)

- Usage of a prompt-based unified data format
  - Minimize influence of variations between tasks
  - Convert tasks into unified multiple-choice generation task

# Are Human Explanations Always Helpful? Towards Objective Evaluation of Human Natural Language Explanations III/V

## Summary

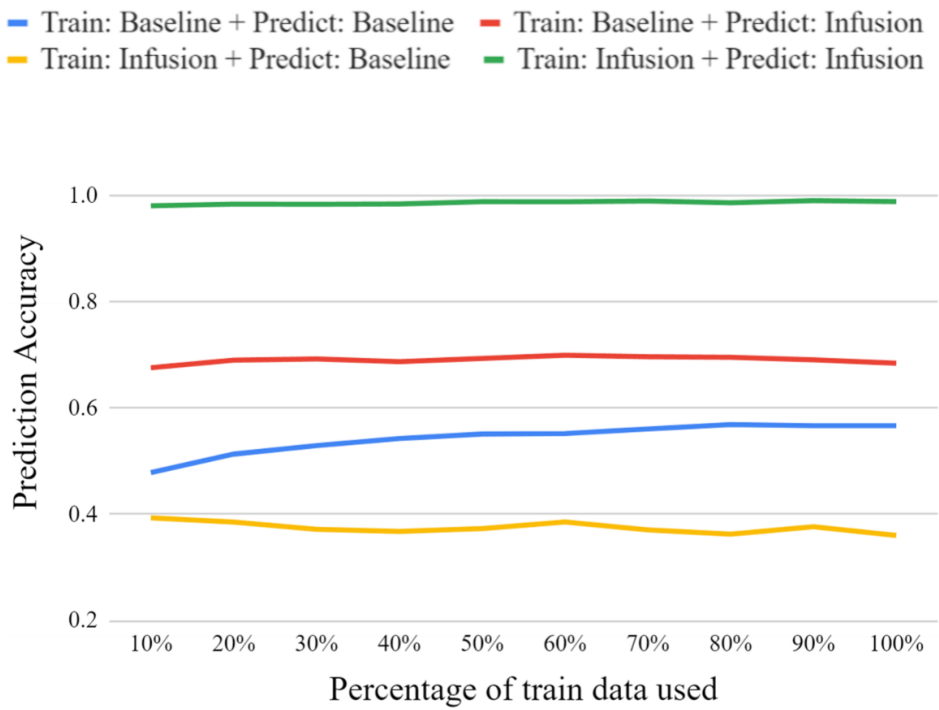



Figure: Prediction accuracies on the ECQA dataset for different amounts of training data and four different setting combinations. Values averaged over three models with different random seeds.

# Are Human Explanations Always Helpful? Towards Objective Evaluation of Human Natural Language Explanations IV/V

## Summary

$$\begin{aligned}\textsc{Treu} =&(Accu(M^{\texttt{Infusion}}_{\texttt{Infusion}}) - Accu(M^{\texttt{Baseline}}_{\texttt{Baseline}}))\\ &+(Accu(M^{\texttt{Infusion}}_{\texttt{Baseline}}) - Accu(M^{\texttt{Baseline}}_{\texttt{Baseline}}))\end{aligned}$$

Figure: The proposed metric; $M^{\textit{predict setting}}_{\textit{finetune setting}}$, where $M$ denotes a model.

- The second summand is the *Simulatability* metric
- *TREU* includes fine-tuning model $M$ using infusion
- *TREU* seems better able to rank explanation quality consistently across multiple datasets and model architectures

# Are Human Explanations Always Helpful? Towards Objective Evaluation of Human Natural Language Explanations V/V

## Similarities

- *TREU* metric is concerned with assessing natural language explanations given for label annotations
- It considers the fine-tuning of models using such explanations

# ActiveAED: A Human in the Loop Improves Annotation Error Detection I/IV

## Summary

- Typical workflow in Annotation Error Detection (AED):
  - Apply AED method, correct flagged errors afterwards
  - No human intervention in the AED step itself
- *ActiveAED*: AED *scoring* method with a human-in-the-loop repeatedly queried for error corrections
  - Requires error scores to rank the erroneous (err.) instances
  - Error corrections are used in its prediction loop
- Based on Area-under-the-Margin metric (AUM)

# ActiveAED: A Human in the Loop Improves Annotation Error Detection II/IV

## Summary

1. Ranking-based AED method finds $k$ most likely annotation errors
2. The (presumed) annotation errors are forwarded to an annotator who provides corrections if nec.
3. The dataset is updated with the corrections

- Repeat until a stopping condition is met, e.g. fraction of errors $<$ thresh.

# ActiveAED: A Human in the Loop Improves Annotation Error Detection III/IV

## Summary

- Annotator has annotation budget of $n$ err. instances
- Either spend the budget on the top-$n$ instances once (SOTA), or repeatedly correct the top-$k$ instances ($k \ll n$) until budget is exhausted (done here)
  - $k := 50$; small enough to be handled in a single annot. session
- Errors defined on sequence level, i.e. sequence is treated as err. if at least one token in it is
- Significantly more compute-intensive than other scoring-based AED methods

# ActiveAED: A Human in the Loop Improves Annotation Error Detection IV/IV

## Similarities

- A human annotator is actively involved in improving the annotations of data
- Does not make use of annot. explanations or LMs during correction

# Explain Yourself! Leveraging Language Models for Commonsense Reasoning I/V

## Summary

- Proposes Commonsense Auto-Generated Explanation framework (CAGE)
  - Improve performance of LMs on commonsense question answering / reasoning (CR) tasks
  - Leverage human explanations in the form of free-text sequences and highlighted span annotations
  - Explanations provide information not present in the input (world knowledge)
  - Highlighting of relevant input portions as additional guidance towards the right answer
  - Both types collected in a new dataset *Common Sense Explanations* (CoS-E)

# Explain Yourself! Leveraging Language Models for Commonsense Reasoning II/V

## Summary

| | |
|---|---|
| Question: | While eating a hamburger with friends, what are people trying to do? |
| Choices: | **have fun**, tasty, or indigestion |
| CoS-E: | Usually a hamburger with friends indicates a good time. |
| Question: | After getting drunk people couldn't understand him,it was because of his what? |
| Choices: | lower standards,**slurred speech**, or falling down |
| CoS-E: | People who are drunk have difficulty speaking. |
| Question: | People do what during their time off from work? |
| Choices: | **take trips**, brow shorter, or become hysterical |
| CoS-E: | People usually do something relaxing, such as taking trips,when they don't need to work. |

Figure: Examples from CoS-E dataset.

# Explain Yourself! Leveraging Language Models for Commonsense Reasoning III/V

## Summary

- CoS-E builds upon Commonsense Question Answering (CQA) dataset, CAGE framework shall generate explanations for CQA
- Two-phased framework:
  1. Provide CQA example (question + answer choices) + corresp. CoS-E explanation to CAGE LM
     - CAGE LM is trained to generate the CoS-E explanation (i.e. an explanation close to it)
     - Correct answer not provided, just the explanation for why it's best ($\rightarrow$ explain-and-then-predict; reasoning)
  2. Use CAGE LM to gen. CAGE explanations for each CQA example
     - Provide them to another CR model (CSRM) which then performs a prediction

# Explain Yourself! Leveraging Language Models for Commonsense Reasoning IV/V

## Summary

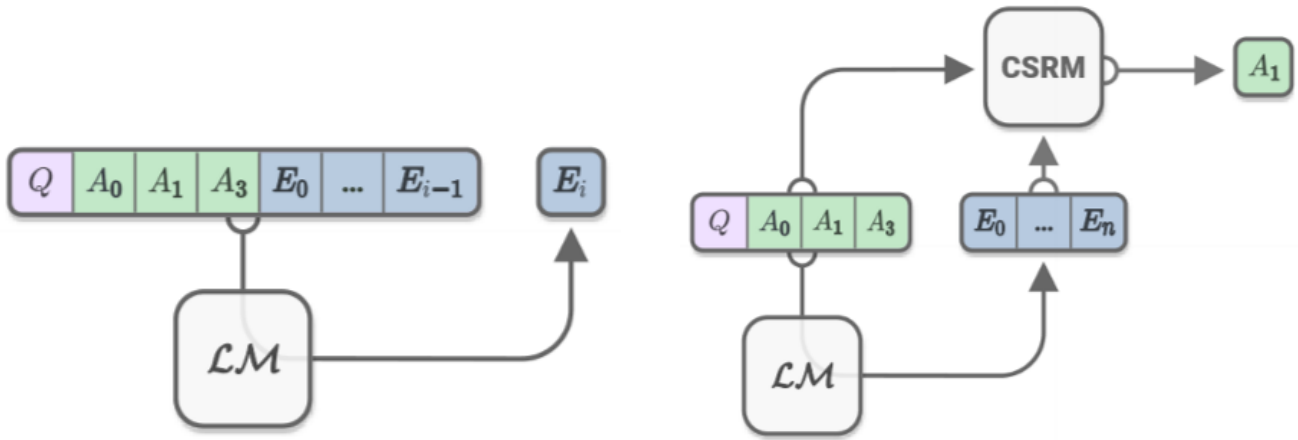



Figure: Two-phased CAGE framework; (1) Train CAGE LM to generate explanations E (token-wise). Conditioned on question Q, answer choices $A_0$, $A_1$, $A_2$ and already generated expl. tokens up to $E_{i-1}$. (2) The trained CAGE LM provides gen. explanation to downstream CSRM.

# Explain Yourself! Leveraging Language Models for Commonsense Reasoning V/V

## Similarities

- Human explanations are used to train / fine-tune an LM
- Auto-gen. explanations are used to aid solving of a prediction task by another LM, similar to annot.
- No human component during inference

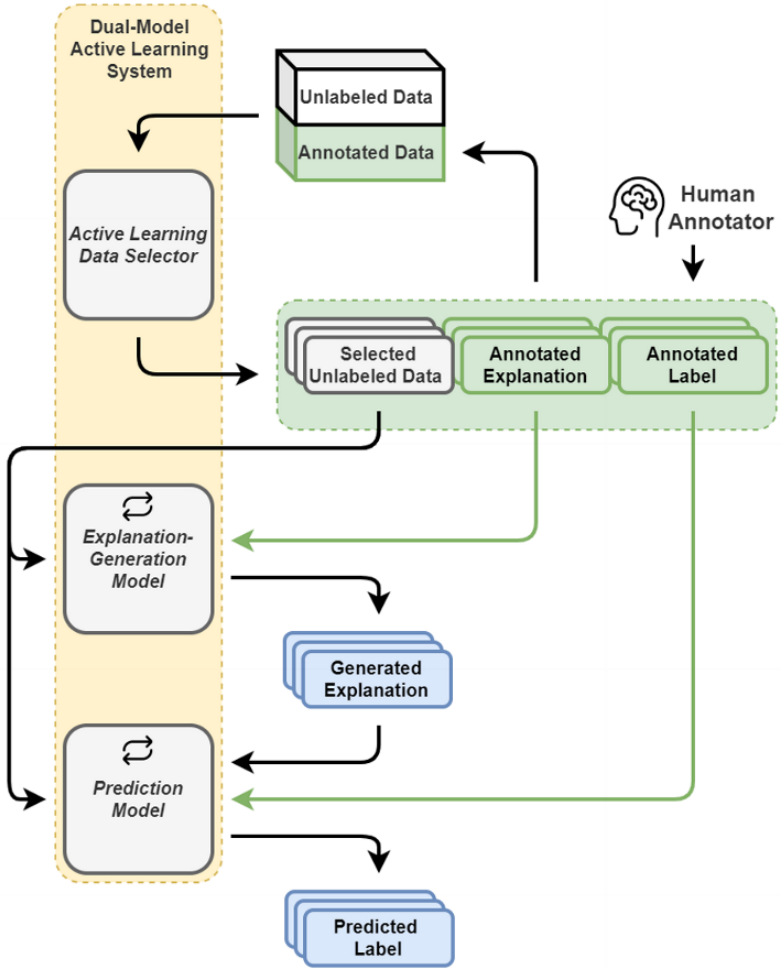
# Beyond Labels: Empowering Human Annotators with Natural Language Explanations through a Novel Active-Learning Architecture I/IV

## Summary

- Active Learning (AL) for data annotation iteratively:
  1. Selects samples from an unlabeled data pool and queries their annot. from human annotators (human-in-the-loop)
  2. Fine-tunes the underlying model with newly annotated data
  3. Evaluates model performance
- Natural-language explanations usually neglected, focus on labelling, though both may be needed
- This work proposes a dual-model AL system arch. that leverages human explanations

# Beyond Labels: Empowering Human Annotators with Natural Language Explanations through a Novel Active-Learning Architecture II/IV

## Summary




- AL data selector chooses few unlabeled examples
- Human provides labels and explanations
- Annot. used to fine-tune Expl.-Gen. model
- Generated expl. and annot. labels used to fine-tune the pred. model
- Humans *review* pred. and gen. expl., start next iter.

# Beyond Labels: Empowering Human Annotators with Natural Language Explanations through a Novel Active-Learning Architecture III/IV

## Summary

- Fine-tuning process teaches the prediction model to rely on explanations

**Explanation-generation Model:**
Training Input **explain:** what is the relationship between *[hypothesis]* and *[premise]* **choice1:** entailment **choice2:** neutral **choice3:** contradiction
Training Target *[human annotated explanations]*
Model Generation *[generated free-form explanation]*

**Prediction Model:**
Training Input **question:** what is the relationship between *[hypothesis]* and *[premise]* **choice1:** entailment **choice2:** neutral **choice3:** contradiction **<sep> because** *[generated free-form explanation]*
Training Target *[human annotated label]*
Model Prediction *[predicted category]*

Figure: Prompt-based input templates for both models

# Beyond Labels: Empowering Human Annotators with Natural Language Explanations through a Novel Active-Learning Architecture IV/IV

## Similarities

- A human-in-the-loop is repeatedly queried to provide guidance for LM in prediction task (similar to ActiveAED)
- Human-annotated explanations used to improve generation of explanations provided by a model

# LEAN-LIFE: A Label-Efficient Annotation Framework Towards Learning from Explanation I/IV

## Summary

- Proposes framework "Label-Efficient AnnotatioN" for sequence labeling and classification tasks
  - Seeks to minimize human annot. efforts
  - Enables learning from explanations for each labeling decision
  - Enhanced supervision / annotation

# LEAN-LIFE: A Label-Efficient Annotation Framework Towards Learning from Explanation II/IV

## Summary

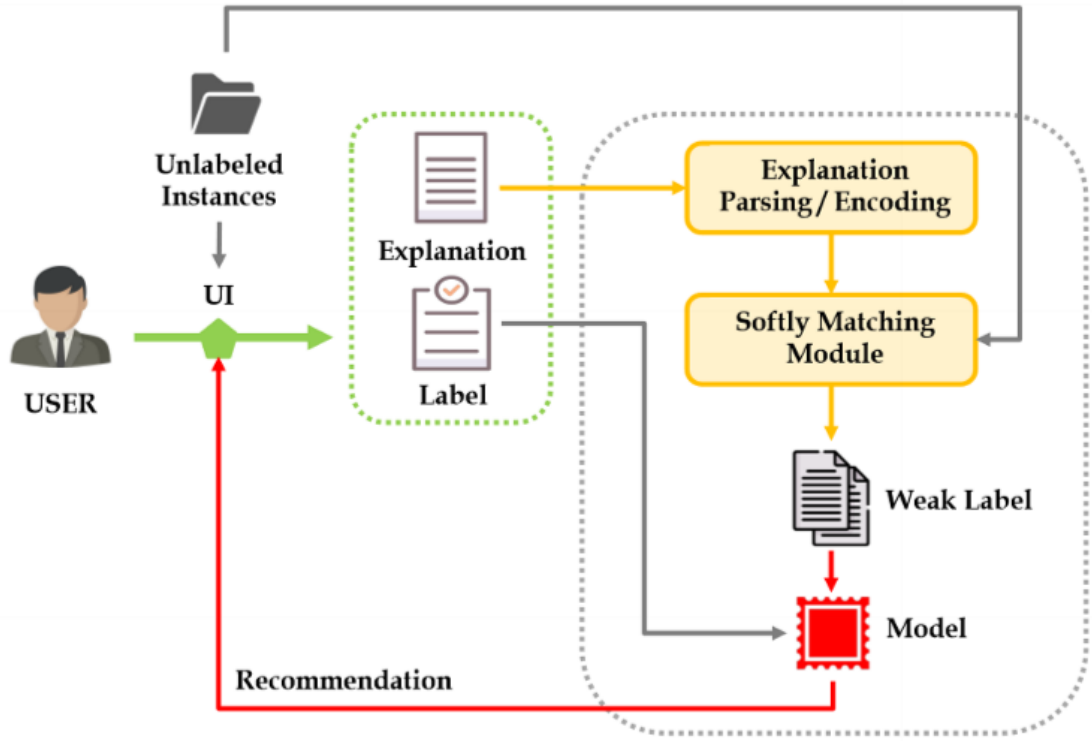



Figure: System architecture; Allow annotators to provide explanations for their decisions (enhanced annot.) either in natural language or by selecting *triggers* at helpful portions. Explanations parsed into labeling *rules* ($\rightarrow$ weakly labeled data), used to provide annot. recommendations.

# LEAN-LIFE: A Label-Efficient Annotation Framework Towards Learning from Explanation III/IV

## Summary

- Soft-matching between trigger representations and unlabeled sentences to generate weakly labeled data
- Natural language explanations parsed into logical form by semantic parser

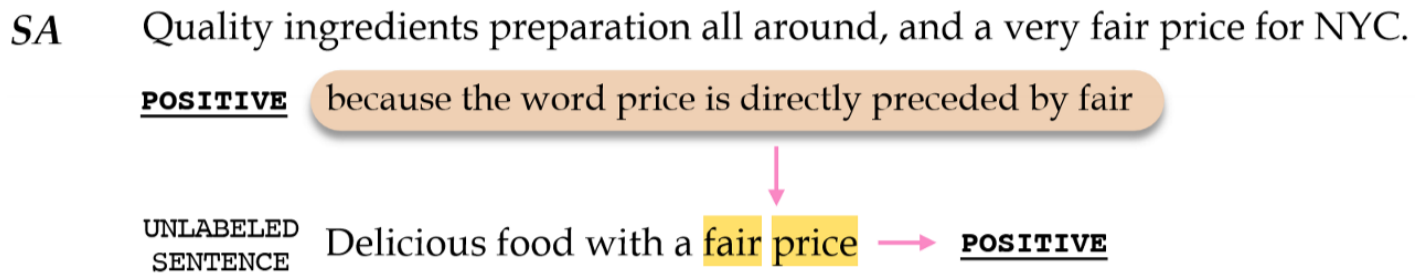



Figure: Example of leveraging labeling explanations on novel data, done here for Sentiment Analysis. The phrase "Delicious food with a fair price" is weakly labeled by the above explanation.

# LEAN-LIFE: A Label-Efficient Annotation Framework Towards Learning from Explanation IV/IV

## Similarities

- A human annotator is aided by a model in labeling data using explanations
- Model is not trained / fine-tuned on generating explanations, but rather on providing recommendations based on previous explanations

# Challenges

- Choice of model
- Source of human-annot. explanations
- Which tasks to tackle (sentiment, offensiveness, etc.)
- Going beyond textual data