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

Datasets

Mark Nagengast Porro

Contents

UKPConvArg1

- Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM [paper, data & code]
- Judge quality of Web arguments, namely convincingness (main goal of argumentation)
- Assignment of "convincingness score" to singular arg. is very subjective (annotator's bias)
  - Instead: Relation classification between arg. pairs
  - A1 more ($>$), less ($<$) or equally ($=$) convincing as A2

UKPConvArg1

- Args. taken from createdebate.com and procon.org
  - Debates w/ $\geq$25 top-level args. of length 10-110 words
  - Sample 25-35 random args. per topic, create $n \cdot (n-1)/2$ pairs
- A *topic* = *prompt* + *stance*
  - *"Should physical education be mandatory in schools? – yes"*
- Each *debate* has two *topics*, one per *stance*
- The args. in a pair cover the same *topic* (i.e. same viewpoint, not combining opposite stances)
- 16.9k arg. pairs, 5 labels + textual labeling reasons (30-140 characters) per pair, 32 topics

UKPConvArg1

- Means of quality control
  - Workers (3,900 total) from the U.S. w/ $\geq$96% acceptance rate
  - Multi-Annotator Competence Estimation (MACE): Estimate true (gold) labels, rank annotators accordingly
    - *threshold* parameter set to 0.95 $\rightarrow$ consider instances w/ entropy among 95% best estimates
    - Reject all assignments of workers that seemingly put in low effort (focus on workers w/ low MACE score); 1161 total
  - Manual checking of reasons
- Three variants: UKPConvArg1-{Full, Strict, Rank}
  - UKPConvArg1-Full: No filtering (apart from MACE pre-filtering)
  - UKPConvArg1-Strict and -Rank: Global filtering using graph construction methods

UKPConvArg1

- Topic-wise: Construct an *argument graph* where args. $\widehat{=}$ nodes and pairs $\widehat{=}$ edges
- Each arg. pair is assigned a weight that quantifies its quality using workers' disagreement and their competence scores
- UKPConvArg1-Strict (11.6k): Discard equal arg. pairs and those that break the DAG properties of the arg. graph
  - The presence of the former causes cycles to break the DAG sooner
- UKPConvArg1-Rank (1k): Rank all args. (nodes) of a topic using PageRank; the higher, the "less convincing"

UKPConvArg1

- While the amount of data is substantial, the explanations are always kept short and their quality/value is lacking at times

  christianity-or-atheism-_atheism.xml:

  - "A1 doesn't go into enough detail."
  - "Neither that good, but a2 is unintelligibly aggressive."
  - "it has a few valid points that I can support."
  - "he brings up good points for his argument"
  - "A1 is too hard to read with the caps."
- Spelling mistakes also a common occurrence

ECQA

- Explanations for CommonsenseQA: New Dataset and Models [paper, data & code]
- Provides explanations for the CommonsenseQA dataset (similar to COS-E)
- Human annotations explain the correct answer choice, refute the incorrect ones

ECQA

Question:
Where is a frisbee in play likely to be?
Answer Choices:
outside park roof tree air

Our Explanation:
Positives Properties
1) A frisbee is a concave plastic disc designed for skimming through the air as an outdoor game.
Negative Properties
1) A frisbee can be outside anytime, even while not in play.
2) A frisbee can be in a park anytime, even while not in play.
3) A frisbee can be on a roof after play.
4) A frisbee can be in a tree after play.
Free-Flow (FF) Explanation
A frisbee is a concave plastic disc designed for skimming through the air as an outdoor game, so while in play it is most likely to be in the air. A frisbee can be outside or in a park anytime, and other options are possible only after play.

ecqa-jsonl, {train, dev, test}_rand_split.jsonl

- For all *(question, correct answer choice, incorrect answer choices)* tuples: Human-annotate positive and negative properties + free-flow explanation
- Around 11k instances
- More basic task, but the explanations are of decent quality and more elaborate
- Only one explanation + set of properties per tuple

e-SNLI

- e-SNLI: Natural Language Inference with Natural Language Explanations [paper, data & code]
- Human annotators given two sentences (premise and hypothesis) + a label (*entailment*, *contradiction* or *neutral*) $\rightarrow$ Provide explanation for label
- Free-text explanations (+ highlighting of relevant passages), generally short and to the point
- 1-3 explanations per pair provided by different annotators
- Also a less subjective task, but w/ multiple explanations per instance and large amounts of data (569k)

e-SNLI

esnli_dev:

- Sentence 1: A white dog with long hair jumps to catch a red and green toy.
- Sentence 2: An animal is jumping to catch an object.
- Label: Entailment
- Explanation 1: A dog is an animal, and a red and green toy is an object
- Explanation 2: White dog is an animal, and toy is object.
- Explanation 3: A dog is an animal.