We Need to Consider Disagreement in Evaluation

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma

Introduction

Evaluation is of paramount importance to Natural Language Processing and Computer Vision

However, today's evaluation practice for virtually all NLP tasks concerned with a fundamental aspect of language interpretation is seriously flawed

Introduction

Predictions are compared against an evaluation set that is assumed to encode a ground truth for the modeling task

The notion of a single correct answer ignores the subjectivity and complexity of many tasks

→ focus on “easy”, low-risk evaluation

What is the background metal structure?

Ms COCO image id 393274, VQA 2.0 question id 393274004

1) trees
2) station
3) awning
4) platform
5) platform
6) platform
7) roof
8) shelter
9) train stop
10) awning

What is the POS tag of 'Anything'?

```
Say Anything with Boyfriend :)
```

Gimpel re-crowsourced dataset

1) PRON
2) ADV
3) NOUN

Gold labels are an idealization, and unreconcilable disagreement is abundant

Similar position

- Plank et al. (2014): Linguistically debatable or plain wrong?
- Jamison and Gurevych (2015), Fornaciari et al. (2021): Noise or additional information?
- Aroyo and Welty (2015): Truth is a lie: Crowd Truth and the seven myths of human annotation
- Uma et al. (2020); Basile (2020): Impact on evaluation of NLP

In contrast

- Bowman and Dahl (2021): study and eliminate biases and artifacts in data
- Beigman Klebanov and Beigman (2009): evaluate on “easy” instances

Sources of disagreement

Individual Differences

Many annotation tasks rely on personal opinions and judgment, despite uniform instructions for annotators
→ For example, in hate speech detection or sentiment analysis

Individual differences can be (partially) explained by cultural and socio-demographic norms and variables, such as age, gender, instruction level, or cultural background

Sources of disagreement

Stimulus Characteristics

Language meaning is ambiguous at several levels: lexical, syntactical, semantic, and others.
→ For example, humour (Raskin, 1985; Poesio, 2020), poetry (Su, 1994) or political discourse (Winkler,2015).

multi-label multi-class vs. multi-class *tout-court*

Sources of disagreement

Context

The same coders could give different answers at different times depending on their state of mind.

Attention slips play a non-negligible role (Beigman Klebanov et al., 2008)

Disagreement in ‘Objective’ Tasks

Disagreement is considered in evaluation of, e.g., machine translation (Papineni et al., 2002) and generation (Lin, 2004).

However it is not in evaluating interpretation:

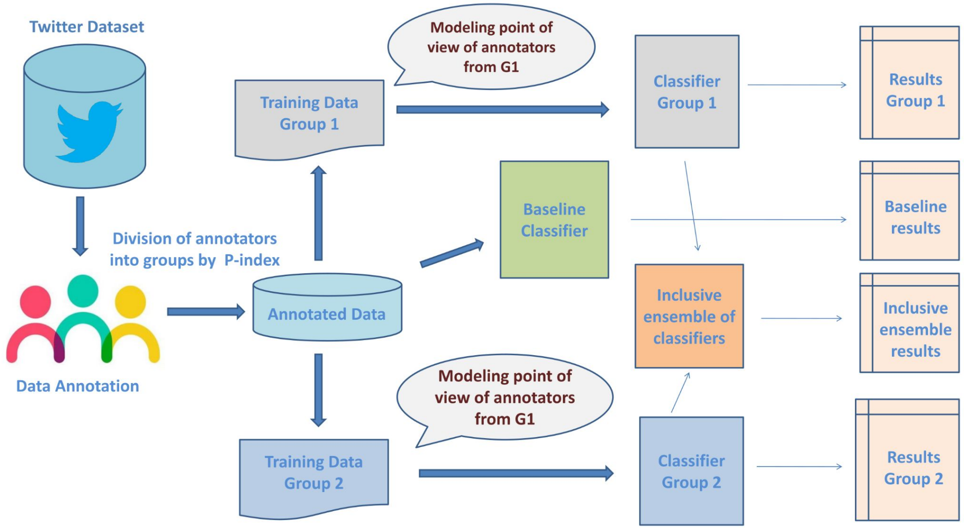
- coreference (Poesio and Artstein, 2005; Recasens et al., 2011)
- part-of-speech tagging (Plank et al., 2014)
- word sense disambiguation (Passonneau et al., 2012)
- semantic role labelling (Dumitrache et al., 2019)

There exist “Inherent Disagreements in Human Textual Inferences” (Pavlick and Kwiatkowski, 2019)

Disagreement on ‘Subjective’ Tasks

Highly subjective tasks such as abusive language and hate speech detection may lead to polarized annotations (Akhtar et al., 2019)

Polarization is a reflection of the cultural background of the annotators and may be exploited to build perspective-aware classifiers (Akhtar et al., 2020)



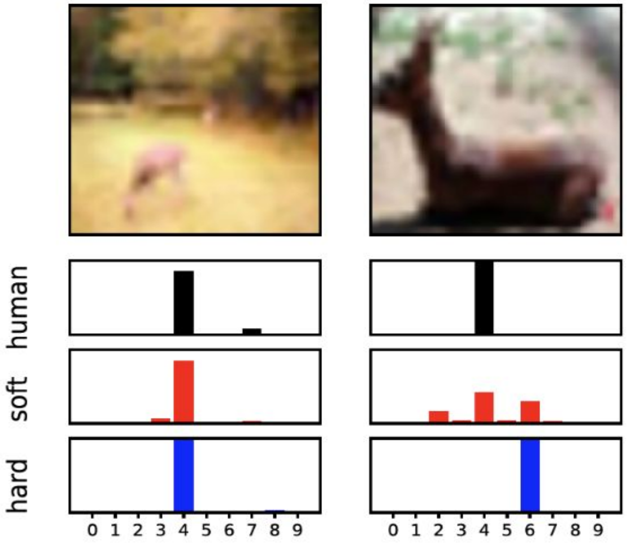
Evaluation in Light of Disagreement

Abandoning the gold standard assumption requires the ability to evaluate a system's output also over instances on which annotators disagree.

Proposals:

- Models produce soft labels evaluated against a full distribution of labels
  → Image classification (Peterson et al., 2019)
  → NLP (Uma et al., 2020; Fornaciari et al. 2021)
- Models produce per-annotator labels evaluated individually
  → "Inclusive" classification (Basile, 2020)

Soft Labels



2: bird, 3: cat, 4: deer, 6: frog

Illustration taken from Peterson et al. (2019)

- Instead of learning from hard labels (e.g. majority human label), learn soft labels (distribution over human labels)
- yields predictions that distribute probability mass more like people, with the same top choice
- left side: same result for soft labels = hard label
- right side: hard label training is hard wrong ("frog"), soft label training gives some probability to correct label ("deer")

SemEval 2021

Task 12: Learning with Disagreements (LeWiDi) (Uma et al.,2021)
A unified testing framework for learning from disagreements in NLP and CV

- Twitter posts annotated with POS tags
- Information Status Classification using the Phrase Detectives corpus
- Humour identification
- Two CV datasets on object identification (LabelMe and CIFAR-10)

Hard evaluation metrics (F1) and soft evaluation metrics (cross-entropy)

Models that account for noise and disagreement have the best (lowest) cross-entropy scores

Evaluation of Highly Subjective Tasks

Aggregated test sets lead to unfair evaluation concerning the multiple perspectives stemming from the annotators' background (Basile, 2020)

Benchmarks for highly subjective tasks should consider the diverging opinions of the annotators throughout the entire evaluation pipeline.

Experiments with models trained on individual annotations
→ impact on explainability, e.g., slurs used by different socio-cultural groups

We argue

against the current prevalent evaluation practice of comparing against a single truth.

→ gross oversimplification of inherently complex matters

We propose

We propose to embrace the complex and subjective nature of task labels.

→ incorporating disagreement leads to better training performance.

→ it can do the same for evaluation (and the datasets already exist).

Subscribe to the Perspectivist Data Manifesto: https://pdai.info