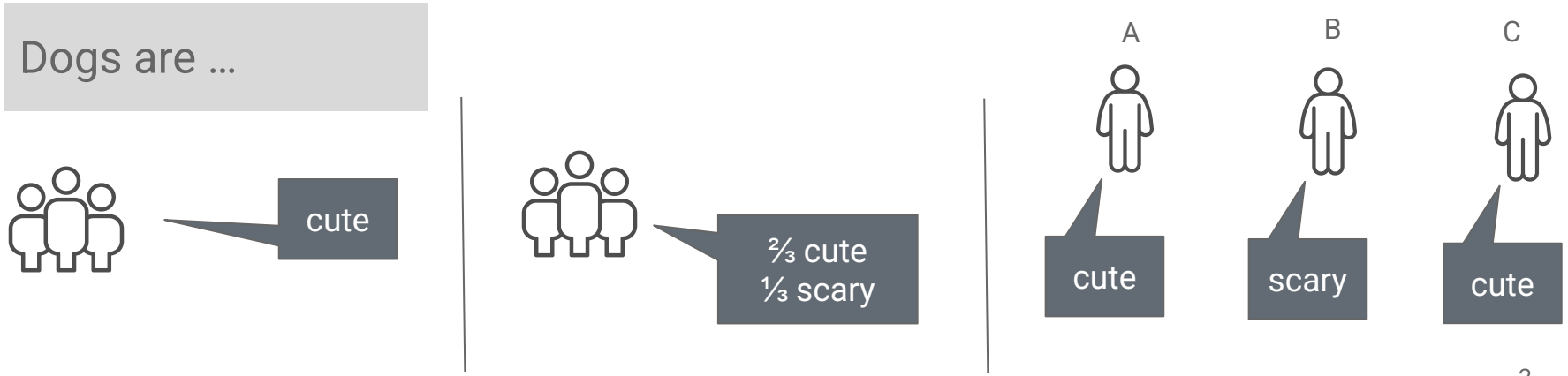
LeWiDi & IDRR

Daniil Ignatev
ESSLLI, August 2026.

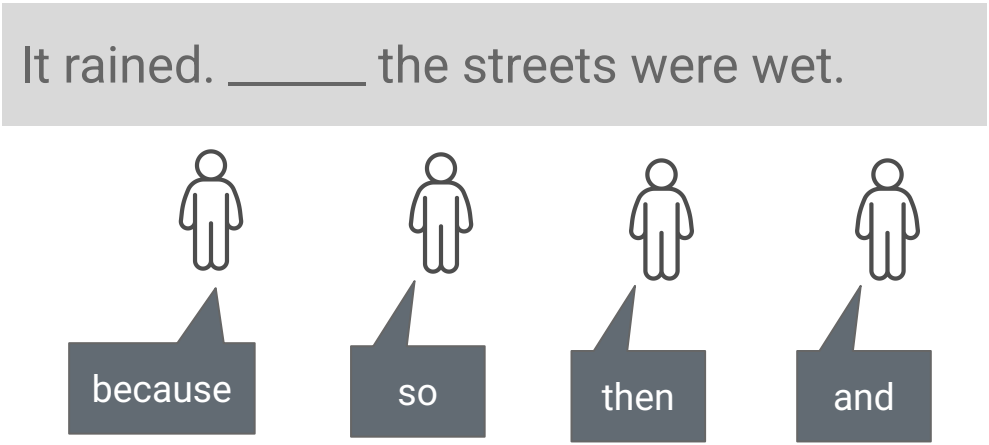
Introduction

- NLP tasks often do not have a single ground truth.
- Human judgments reflect diverse perspectives, backgrounds and reasoning styles.
- Need models that handle variation.



Implicit Discourse Relation Recognition (IDRR)

- IDRR involves identifying relations without explicit markers
- High ambiguity and cognitive complexity
- Disagreement ≠ error



P1: Human Label Variation in Implicit Discourse Relation Recognition

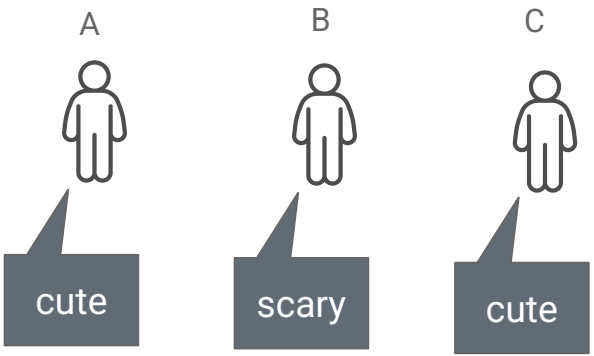
Frances Yung, Daniil Ignatev,
Merel Scholman, Massimo Poesio
and Vera Demberg
LREC 2026

Traditional Approach

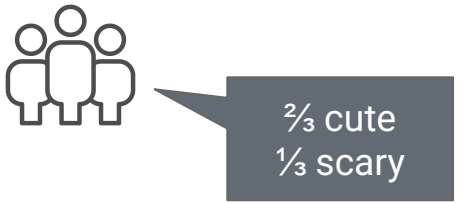
- Single-label classification
  - Penn Discourse Treebank (PDTB) multi-labels: 5%
  - Agreement between 2 experts = 44% (Hoek+ 2021)
- Ignores disagreement and interpretive diversity
- Limits model performance: SOTA 64% Accuracy (Wang+ 2025)

Need for New Models

- Perspectivist models: learn the perspectives of individual annotators



- Label-distribution models: predict probability distributions



Perspectivist Models

- Capture individual labeling perspectives
- Designed to handle subjective variation:
  - Hate speech / toxicity detection
  - Sentiment analysis

Methods: e.g. Separate annotator heads or embeddings

Davani+ 2022; Deng+ 2023

Research Gap

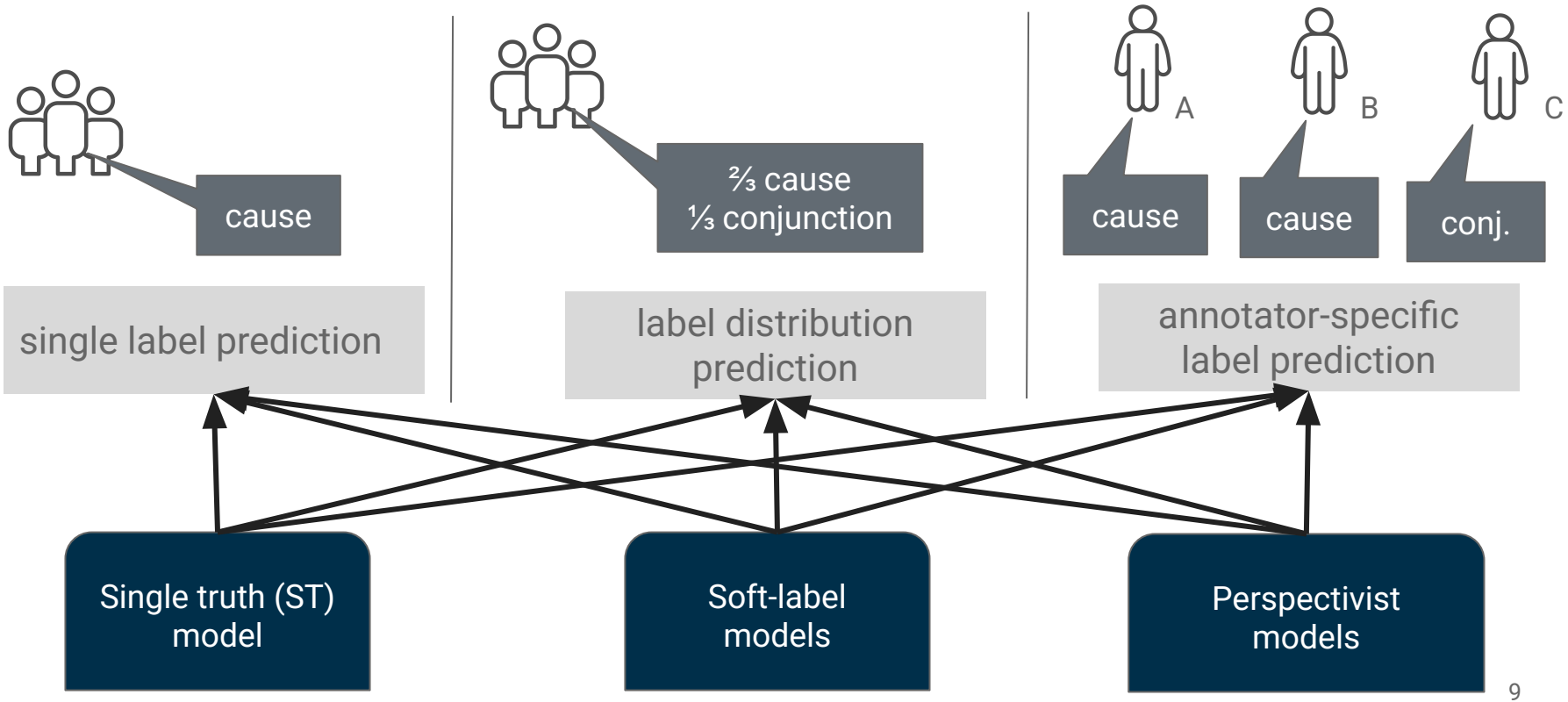
- Perspectivist models not tested on IDRR
- IDRR: Highly ambiguous and cognitively challenging
- Cognitive-complexity-driven (instead of subjectivity-driven) disagreement

The lights went out. _____ People started talking.

Lights going out implies some show has finished, so people talk. → result

Two independent events; just more information. → conjunction

Experiment

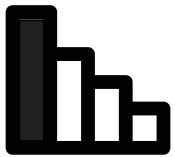

cause
⅔ cause
⅓ conjunction
A
B
C
cause
cause
conj.
single label prediction
label distribution prediction
annotator-specific label prediction
Single truth (ST) model
Soft-label models
Perspectivist models

Dataset

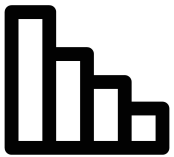
- **DiscoGeM 1.5 corpus**: multi-genre, English, multi-annotator (Scholman+ 2022, Yung & Demberg 2025)
- Derived from PDTB-style implicit DRs
  - Sense hierarchy: Level-1 (5 classes), Level-2 (17 classes)
- 10 annotations per instance
- Connective insertion method:
  - “because” ⟶ cause
  - “however” ⟶ concession
  - “or” ⟶ disjunction

Baseline Single-Truth Model (ST)

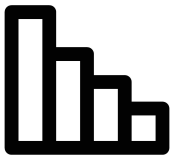
- RoBERTa-base trained on majority labels
  - Training objective: most probable class



Soft Label Models



1. **Multi-label model** with Binary cross-entropy (BCE) loss
   - Trained on labels with >=2 votes out of 10
   - Training objective: predict multiple labels



2. **Label-distribution** model with KL-divergence loss
   - Trained on raw label distributions
   - Training objective: predict label distributions

Perspectivist Models

1. Multi-task (MT): one classification head for each annotator
   (Davani+ 2022)
2. Annotator Embedding (AE): trained annotator vectors
   (Deng+ 2023)
   - Annotator IDs passed as features during training;
   - Objective: predict annotator-specific labels;

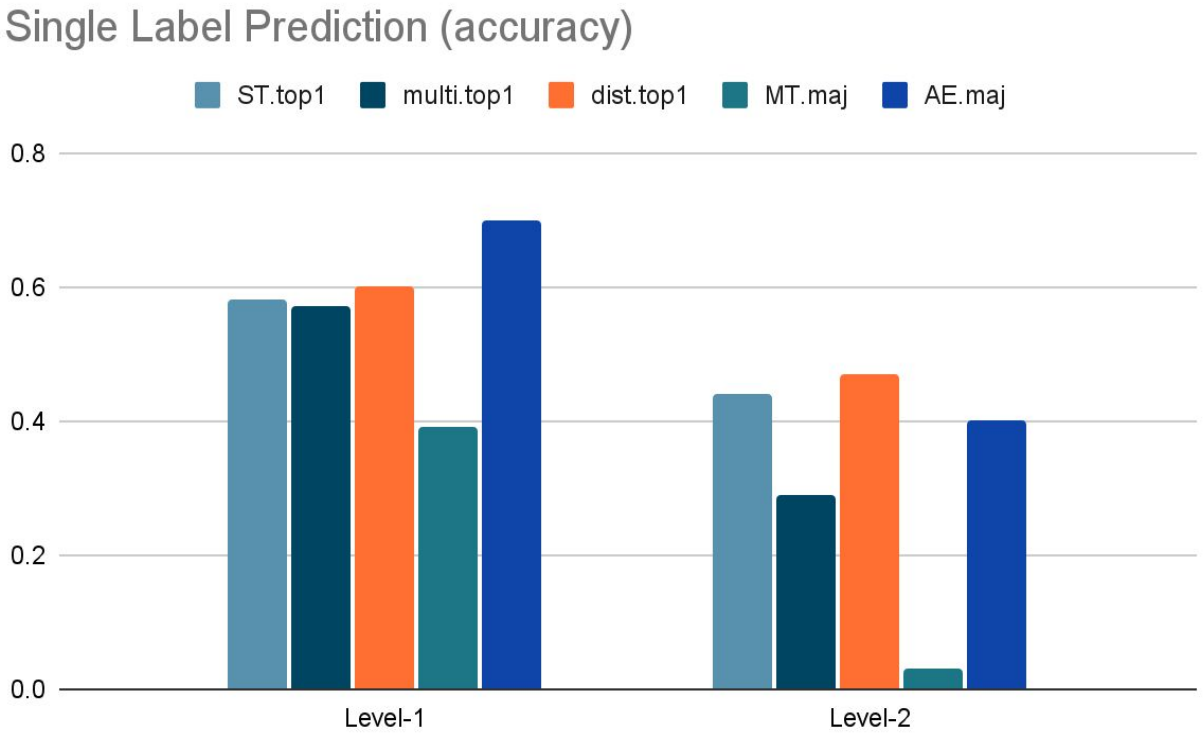
Annotator-specific predictions

(A, cause),
(B, cause),
(C, conjunction) …..

Single-label Prediction

Model	Default output	Task output
Single Truth (ST)	single label	single label
Multi-label	multiple label	**single label with highest probability**
Distribution (dist)	label distribution	
Multi-task	annotator-label pairs	**majority label after grouping by instance**
Annotator Embedding (AE)		

Single-label Prediction

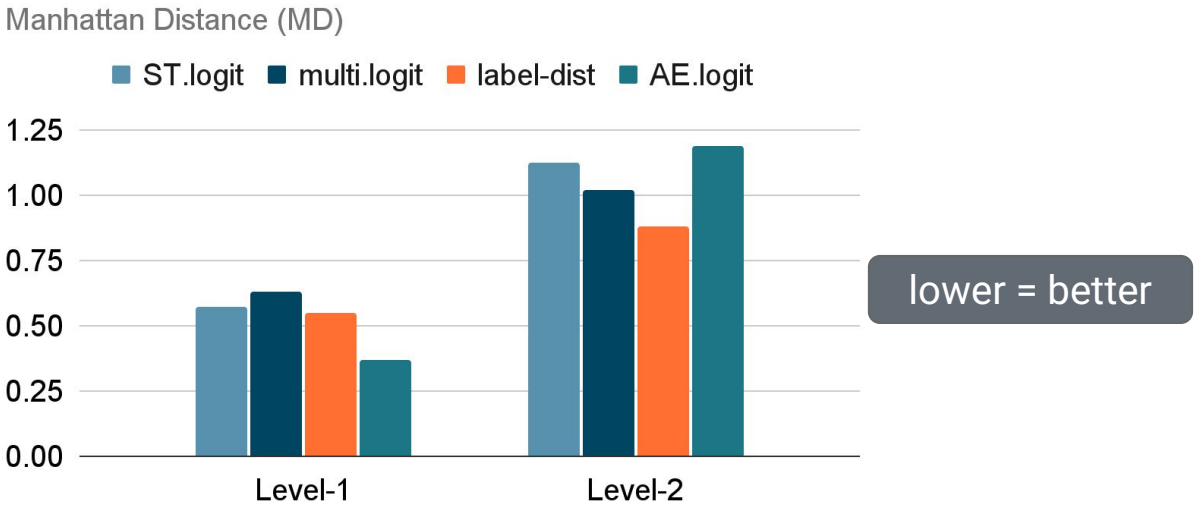


- Label-distribution (dist.top1) improves over ST baseline
- Annotator Embedding model best on Level-1
- Higher label granularity reduces accuracy (Multi-Task model drops severely)

Label Distribution Prediction

Model	Default output	Task output
Single Truth (ST)	single label	**logits of prediction**
Multi-label	multiple label	
Distribution (dist)	label distribution	label distribution
Multi-task	annotator-label pairs	**averaged annotator-specific logits**
Annotator Embedding (AE)		

Label Distribution Prediction



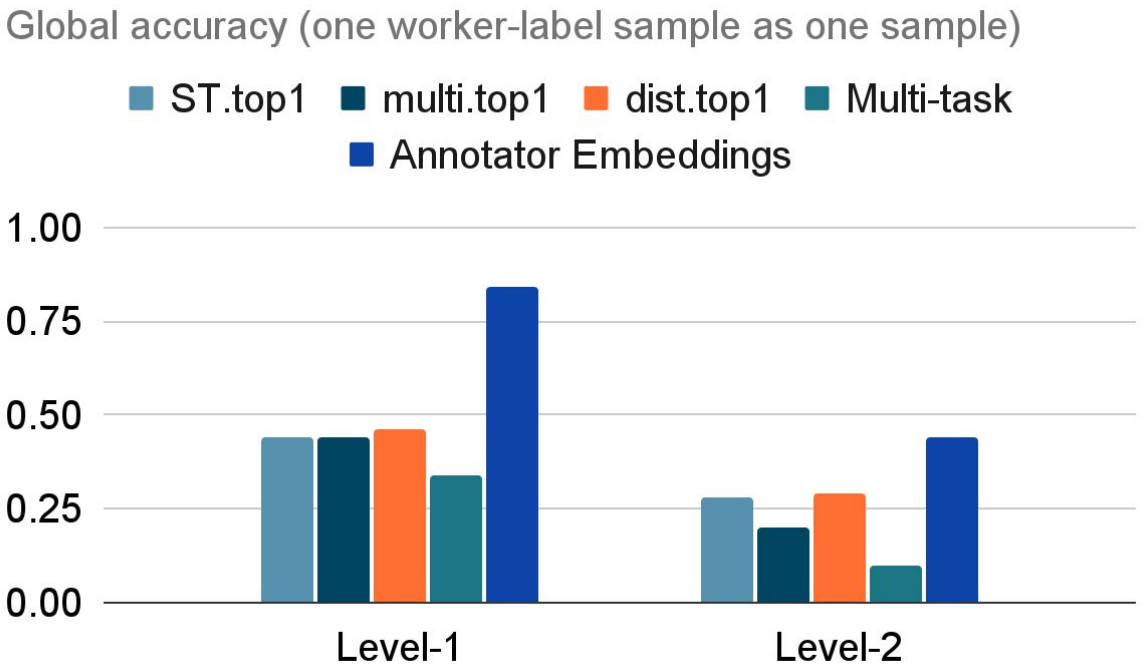
- Label-distribution model best overall

  ⟶ Soft-label training captures human variation
- Perspectivist models (AE.logit) good only at coarse levels

Annotator-Specific Prediction

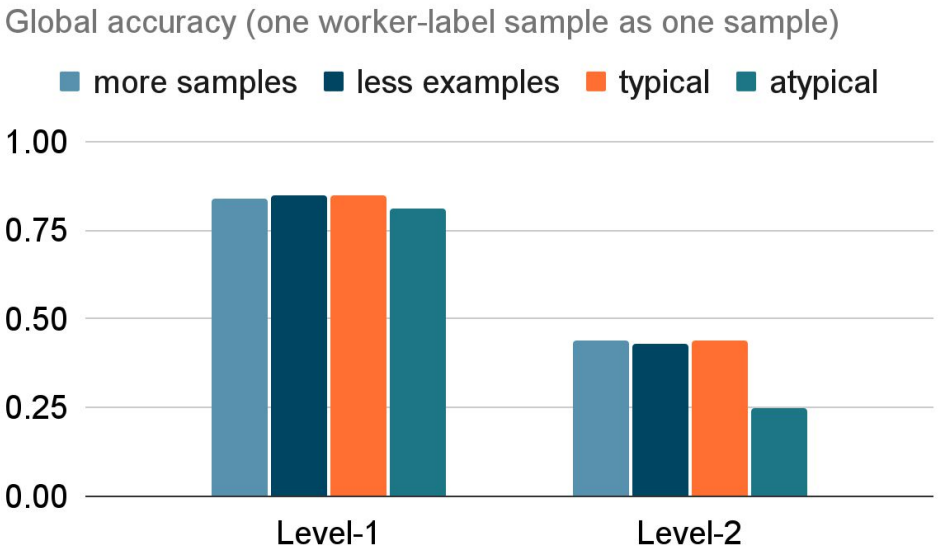
Model	Default output	Task output
Single Truth (ST)	single label	**same label for all annotators**
Multi-label	multiple label	**same top1 label for all annotators**
Distribution (dist)	label distribution	
Multi-task	annotator-label pairs	annotator-label pairs
Annotator Embedding (AE)		

Annotator-Specific Prediction



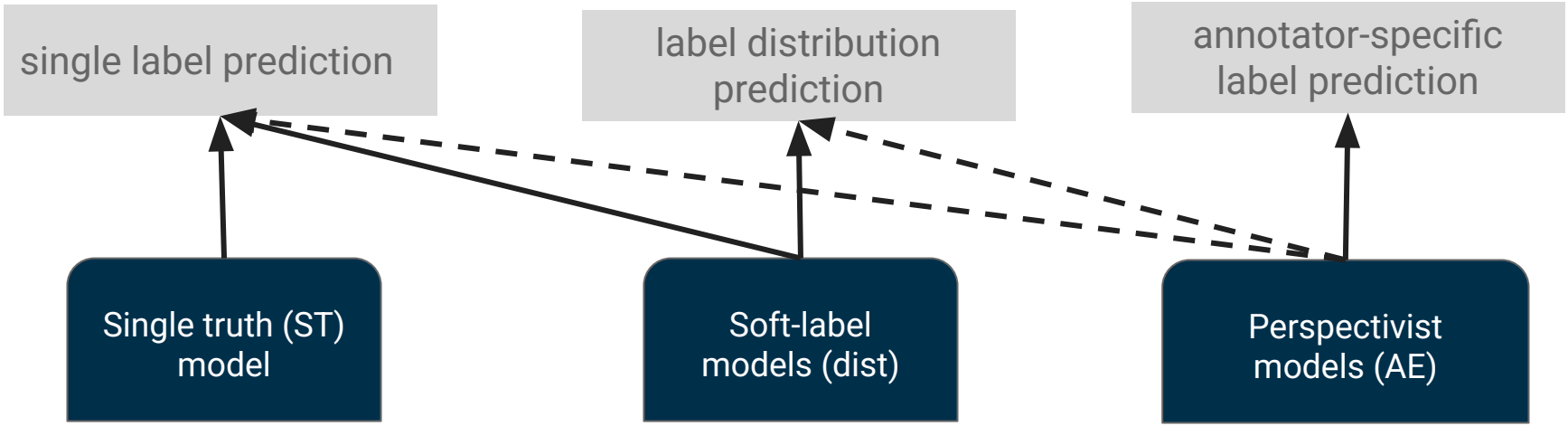
- Annotator Embedding model outperforms others
- Multi-task model fails to generalize (too many heads)

Annotator-Specific Prediction per annotator group



- Compare AE model performance on different subsets of annotators:
  - more/less per-worker training samples
    - no major difference
    - dataset is large enough
  - typical/atypical with respect to the majority
    - model performs **worse on atypical** workers

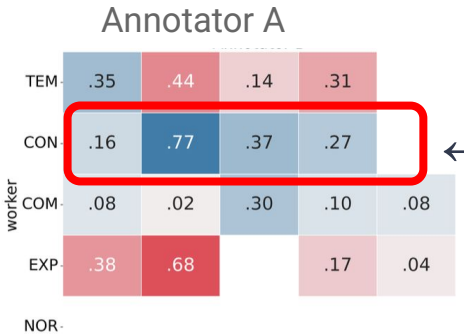
Key Findings



- Distribution learning can generalize to majority prediction
- Perspectivist models succeed only at Level-1
- Ambiguity limits performance (20% drop between Level-1 and Level-2)

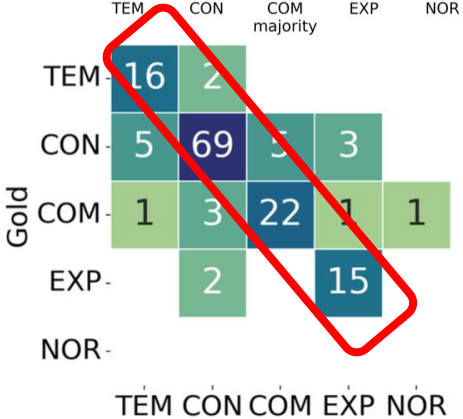
Perspective model captures some biases, misses others

worker vs majority nPMI



← bias to concession

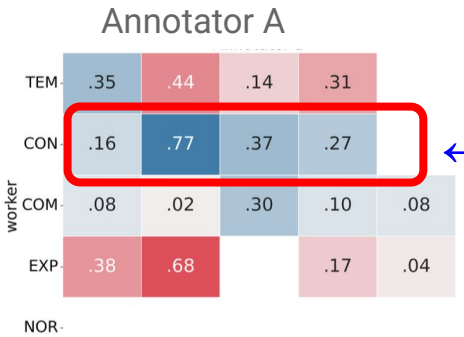
gold vs prediction confusion matrix



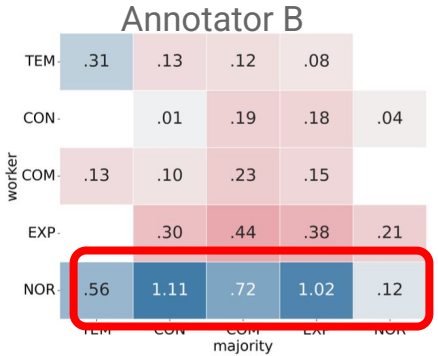
← model captures the bias

Perspective model captures some biases, misses others

worker vs majority nPMI

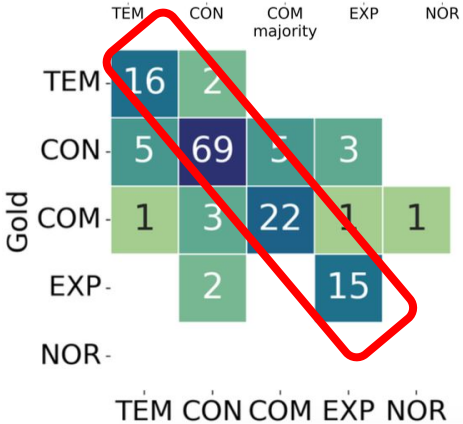


← bias to concession



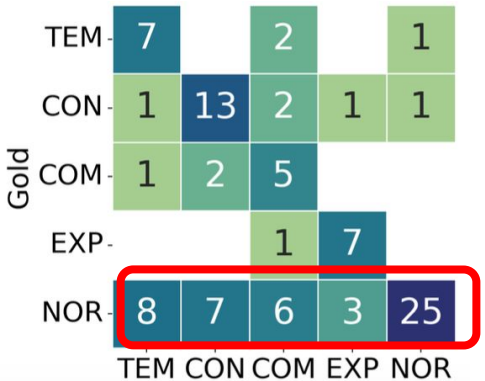
bias to no-relation →

gold vs prediction confusion matrix



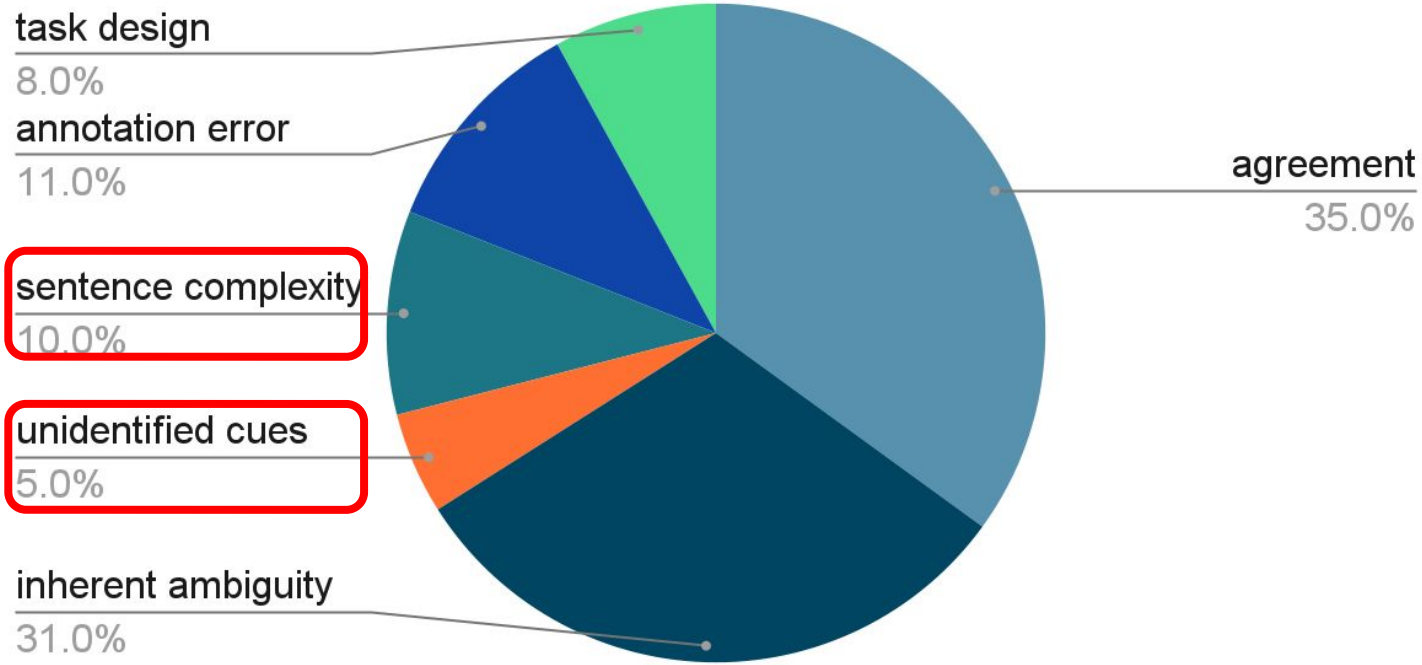
← model captures the bias

model underpredicts no-relation →



Sources of variation

- Analyzed 3 workers' annotations on 100 items.



Cognitively difficult cases lead to inconsistent behaviour.

Cognitively demanding cases

- sentence complexity

While it had devastated me and stopped me in my tracks, for Maxime, it had broken down barriers, released him from a straitjacket, and left him free to write his own story. _____**After** what happened, I was never the same.

“after” read as connecting two sentences → asynchronous

“after” correctly parsed as intra-sentential → contrast

Cognitively demanding cases

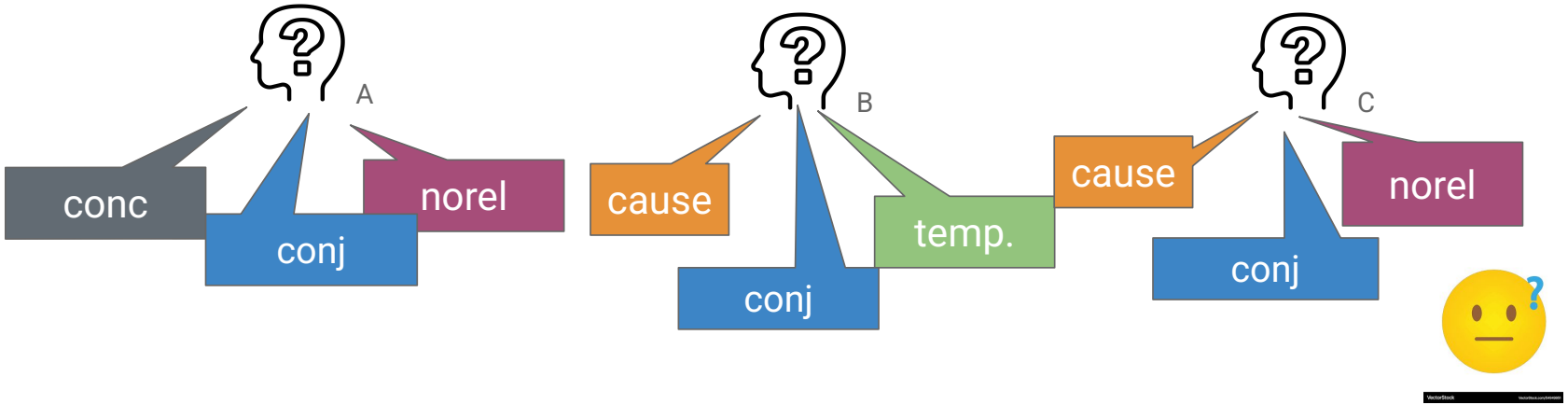
- sensitivity to discourse cues

...I lived in fear and in a constant state of mental anguish that led me to disastrously fail my final exams._____**By the summer of ’93**, I had already left the Côte d’ Azur for Paris, where ... I enrolled in a second-rate business school.

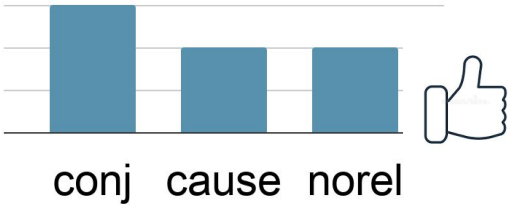
sensitive to temporal cues —> asynchronous

not sensitive to temporal cues → causal

Conclusions



- IDRR disagreement reflects ambiguity and cognitive variability.
- The ambiguity is **challenging for perspectivist modelling.**
- **Label-distribution** modelling is more robust.



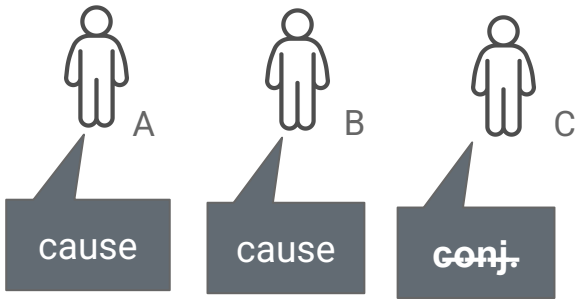
P2: Dataset Cartography for IDRR: Promises and Pitfalls

Daniil Ignatev,
Denis Paperno,
and Massimo Poesio
CODI-CRAC 2026

Utrecht University

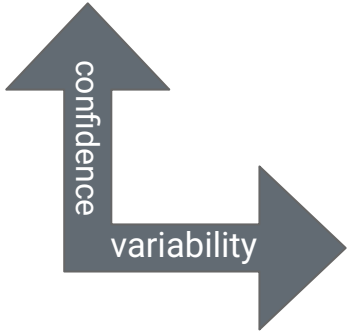
Motivation

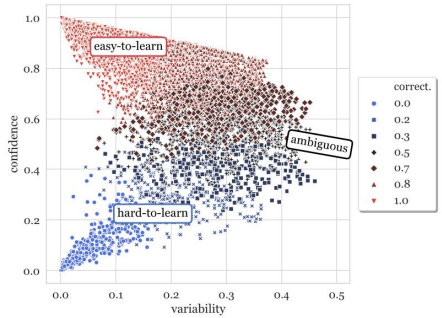
- Some of the crowdsourced labels result from lapses of attention.
- Existing algorithms fail to fully distinguish the (un)reliable workers / items. Example: NUTMEG (Ivey et al. 2025).
- Robust solution needed to filter out implausible interpretations.



Dataset Cartography, p. 1

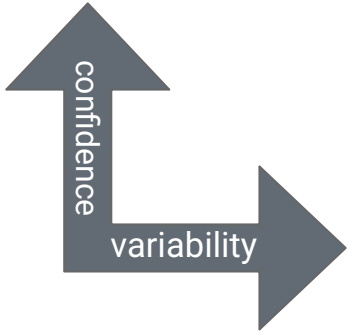
- Dataset cartography (Swayamdipta et al., 2020) widely used for model and data diagnostics.
- Training a supervised classifier, we extract relevant parameters ("training dynamics") at each training epoch.
- Predicted probability of the correct class across the epochs, **P_corr**, characterizes each training data point; e.g., stable vs flipping.

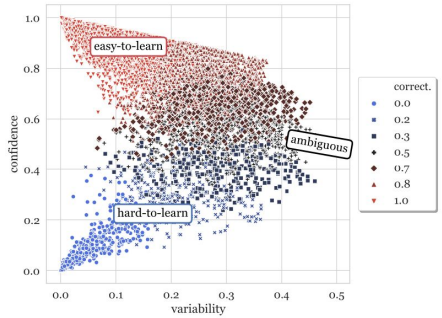




Dataset Cartography, p. 2

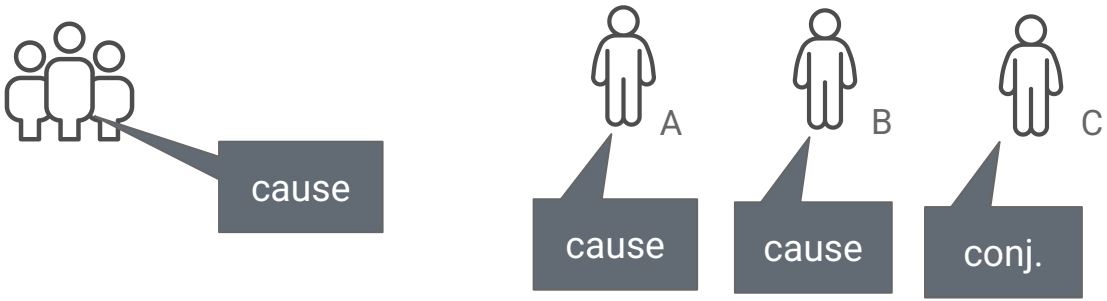
- Main parameters: confidence = mean **P_corr**, variability = standard deviation of **P_corr**.
- High/low confidence and variability = **easy/hard to learn**.
- The hard/easy distinction may be able to separate errors from plausible annotations (Klie et al., 2023; Weber-Genzel et al., 2024); primary evidence from NLI; also coincides with IAA.



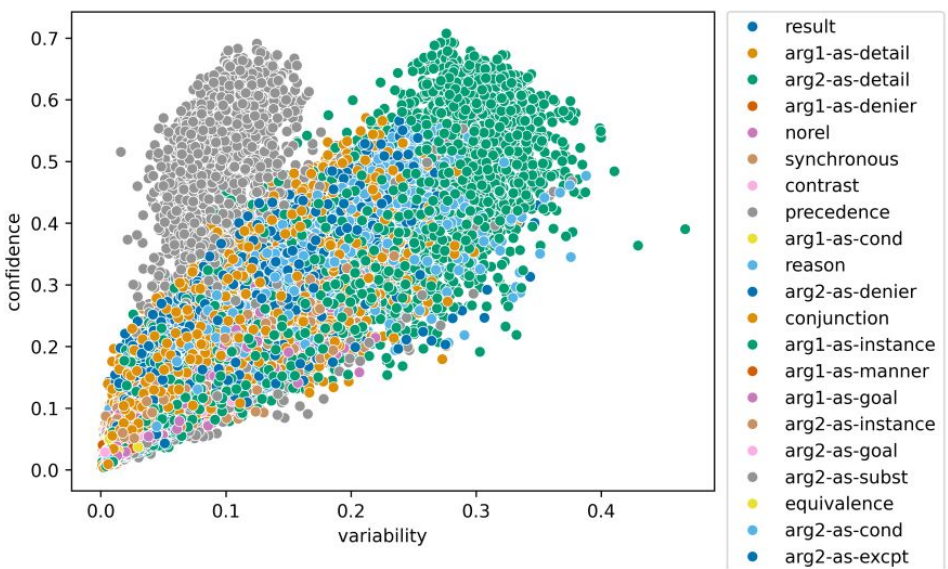


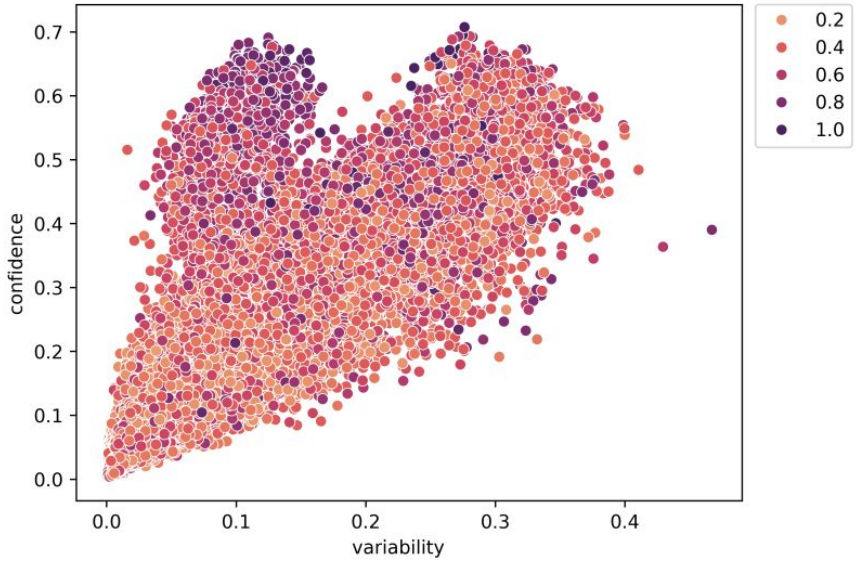
Setup

- Model 1: "normal" single-label classifier, **SG**
  - trained on majority-aggregated labels, **maj** (1).
- Model 2: perspective-aware classifier by Deng et al. (2023), **AE**
  - trained on raw labels and annotator IDs, **raw** (2).
- DeBERTa-v3-base as the base model. Data: DiscoGeM 1.5



Data Maps (lvl3, AE)





Hard-to-learn

- Not feasible as a filtering criterion due to label imbalance. Sparsely represented labels are also hard-to-learn, cf. the odds ratio. The expected trend, **hard-to-learn = low agreement**, is **not** confirmed.
- Sparse classes harder to learn: at granular levels (1), for annotator-aware training, AE (2).

	Level-1	**Level-2**	**Level-3**
LowAgree.OR.SG	1.75	1.48	1.51
RareLabel.OR.SG	1.74	0.54	3.75
LowAgree.OR.AE	1.13	1.31	1.46
RareLabel.OR.AE	7.47	17.21	25.8

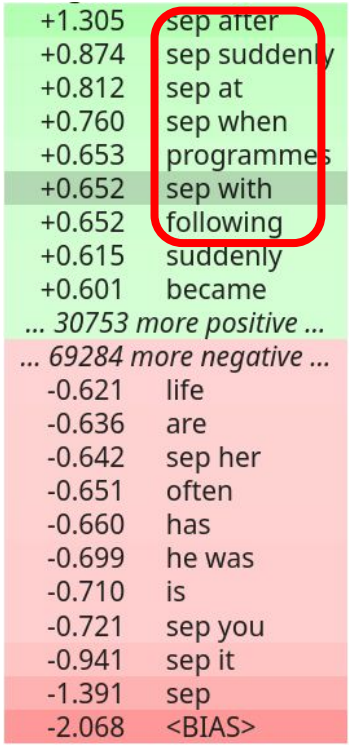
Easy-to-learn

- A set of "overspecified" data with connectives at the start of Arg2 consistently stands out as easy. ARG1*: [It] left him free to write his own story.* ARG2*: **After** what happened, I was never the same.*
- Why still implicit: the connective is intra-Arg2, does not apply to the Arg1-Arg2 relation.
- The set coincides with high agreement (1).
- The set coincides with temporal relations (2); particularly, precedence.
- Likely shows the annotators taking hints.

Lexical Probe, p. 1

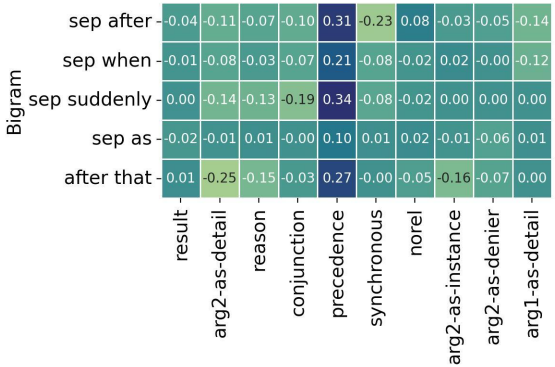
Metric	Level-1	Level-2	Level-3
Acc.Maj	0.86 ± 0.02	0.90 ± 0.02	0.90 ± 0.01
F1.Maj	0.62 ± 0.03	0.60 ± 0.06	0.61 ± 0.02
Acc.Raw	0.87 ± 0.00	0.93 ± 0.00	0.93 ± 0.00
F1.Raw	0.67 ± 0.00	0.71 ± 0.01	0.68 ± 0.02

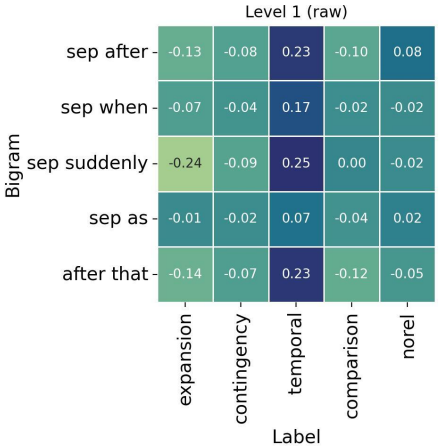
- Confirmation via lexical probe: predicting temporal labels from bigrams.
- How to make sensitive to the Arg1/Arg2 boundary? Use a SEP word
- Results averaged across 5-fold cross-validation.
- Important positive features are lexical hints.

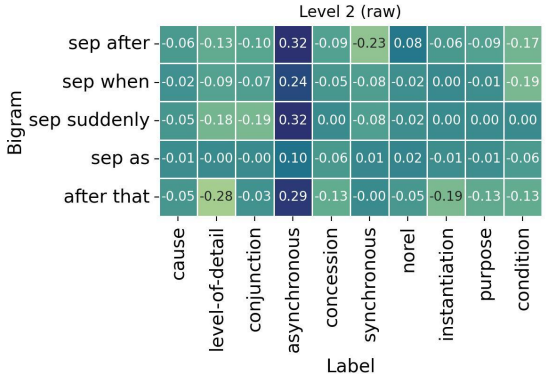


Lexical Probe, p. 2

- Normalized PMI between the detected cues and top-10 labels at each level: association with "temporal" / "precedence".







Takeaways

- Promises: for IDRR, DC can identify "overspecified" data where the crowd workers converge on a hinted label, sometimes erroneously.
- Could extrapolate to other tasks where hints matter.
- Pitfalls: DC cannot reliably identify noisy labels when the label taxonomy is highly granular.
- Implications: "overspecified" implicit relations in existing datasets need to be clearly marked to prevent conflated model evaluation.
- For ambiguous tasks, easy-to-learn data may be more straightforward to diagnose.

Takeaways (global)

- Our position is that most labels in crowdsourced IDRR are valid readings.
- These labels can be effectively used to obtain more robust predictive models.
- IDRR data are non-homogenous: human label variation coexists with problematic data instances, more reliable ways of separating these groups have yet to emerge.