IT UNIVERSITY OF COPENHAGEN

What to do about Human Disagreement in Natural Language Processing?

@barbara_plank

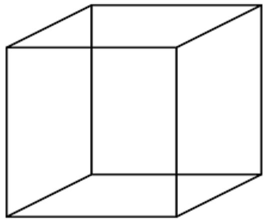
DALI end-of-project workshop

September 24, 2021

Disagreement in human annotation is ubiquitous

Side benefit of annotation - fortuitous data:

Disagreement as a source of information?



Roadmap

1. Data: Is disagreement random noise?
2. Modelling: How can we leverage disagreement?
3. Evaluation: How to evaluate in light of disagreement?

Selected examples

Act I: Data

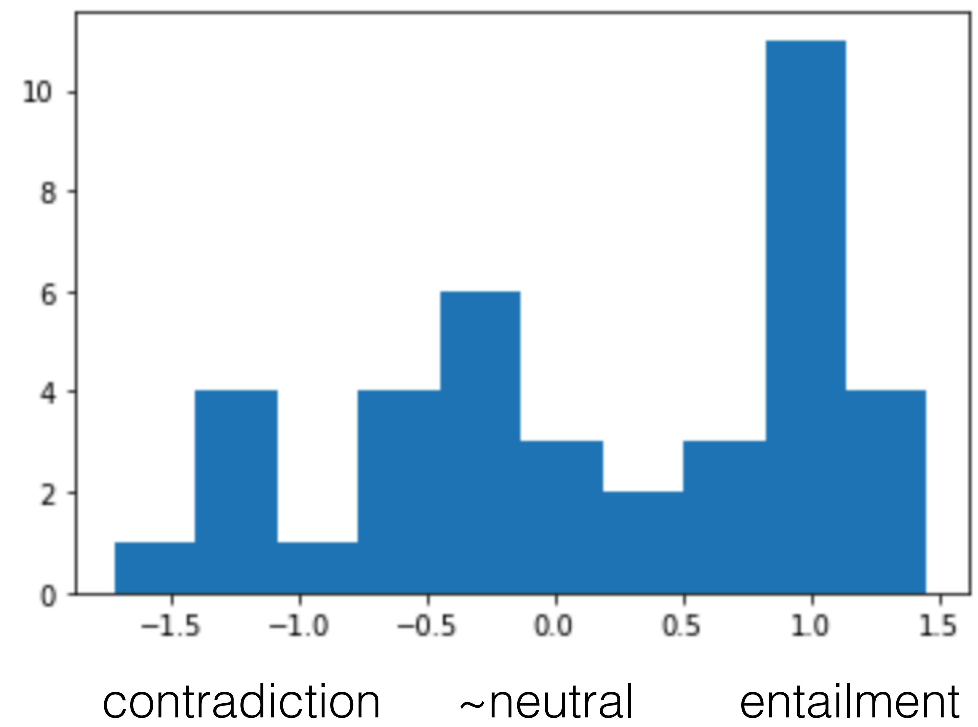
Medical Relations Extraction (MRE)

relation, count
ASSOCIATED_WITH 4
SYMPTOM 3
CAUSES 3
PREVENTS 1
SIDE_EFFECT 1
MANIFESTATION 1
PART_OF 1
DIAGNOSE_BY_TEST_OR_DRUG 1
OTHER 1

These data suggest that subclinical RIBOFLAVIN DEFICIENCY may occur in adolescents and that deficiency may be related to dietary intake of RIBOFLAVIN

Data and example from Aroyo & Welty (2015) and Palomaki et al., (2018)

Recognising Textual Entailment (RTE)



Premise *p*: Amanda carried the package from home .
Hypothesis *h*: Amanda moved .

Does p->h?
original-dataset-label: entailed

there are linguistically hard cases, even for POS tagging

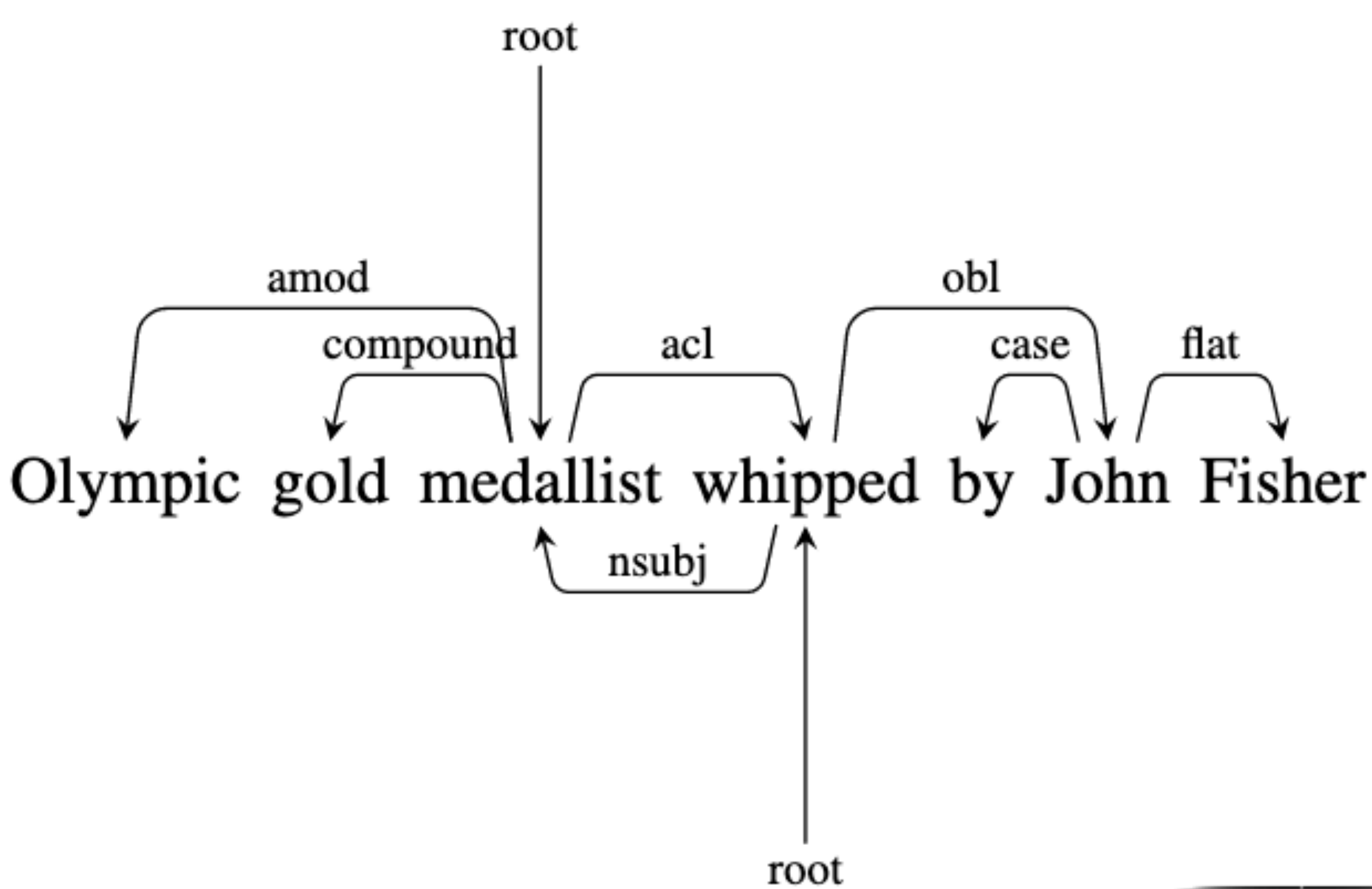
e.g. Manning (2011). *Part-of-Speech tagging from 97% to 100%. Is It Time for Some Linguistics?*

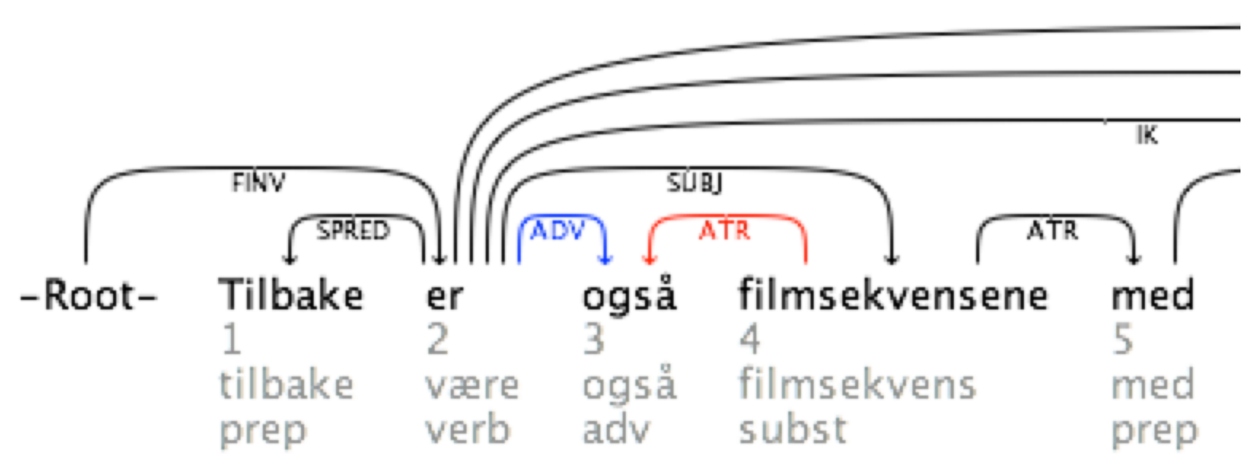
Part-of-Speech (POS)

VERB	**NOUN**	ADP	NOUN	SYM
VERB	**PRON**	ADP	NOUN	SYM
VERB	**ADV**	ADP	NOUN	SYM
Say	Anything	with	boyfriend	:)

Example from Twitter crowdsourced dataset (Gimpel et al., 2012; Hovy et al., 2015)

Dependency Parsing





Example from Liu et al., 2018

Example from Norwegian data in Martinez-Alonso et al. 2015

Understanding Indirect Questions

Q: Hey. Everything ok?
A: I'm just mad at my agent

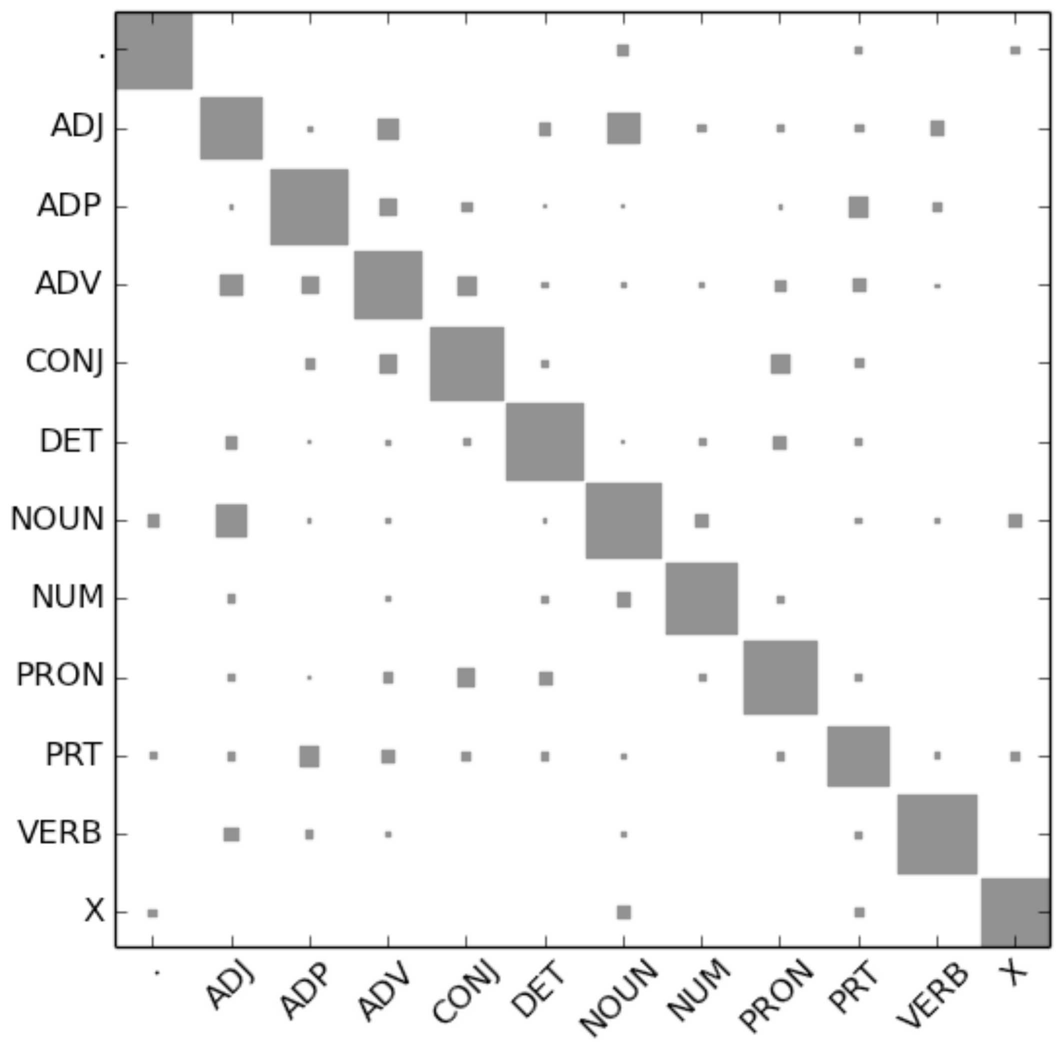
Yes
No
Yes, subject to some condition
Neither Yes nor no
Other
N/A

Damgaard, Toborek, Eriksen & Plank (2021) To appear in CODI, EMNLP 2021 workshop
Louis et al. (2020) EMNLP

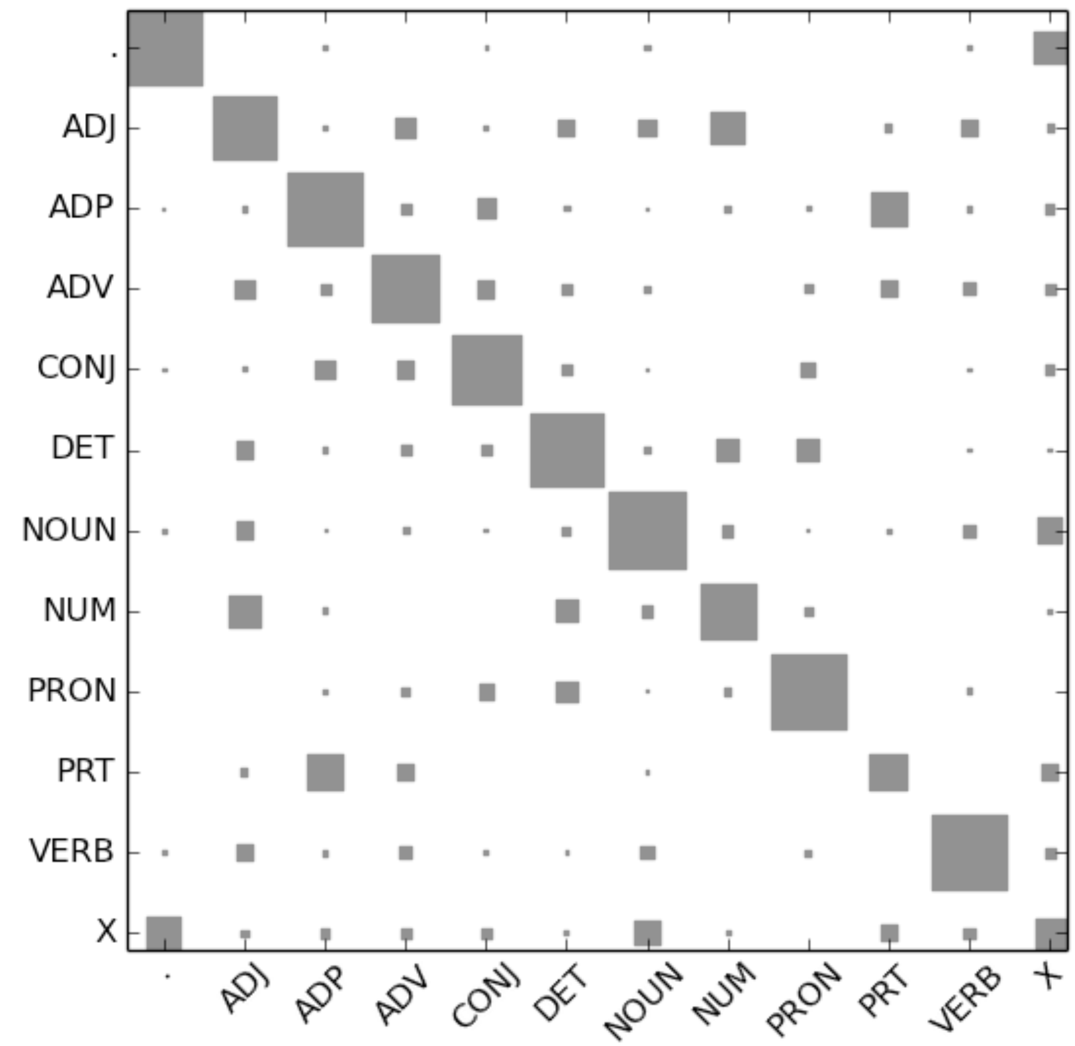
Are disagreements randomly distributed?

... and can we estimate disagreements from small samples?

(Plank et al., 2014)

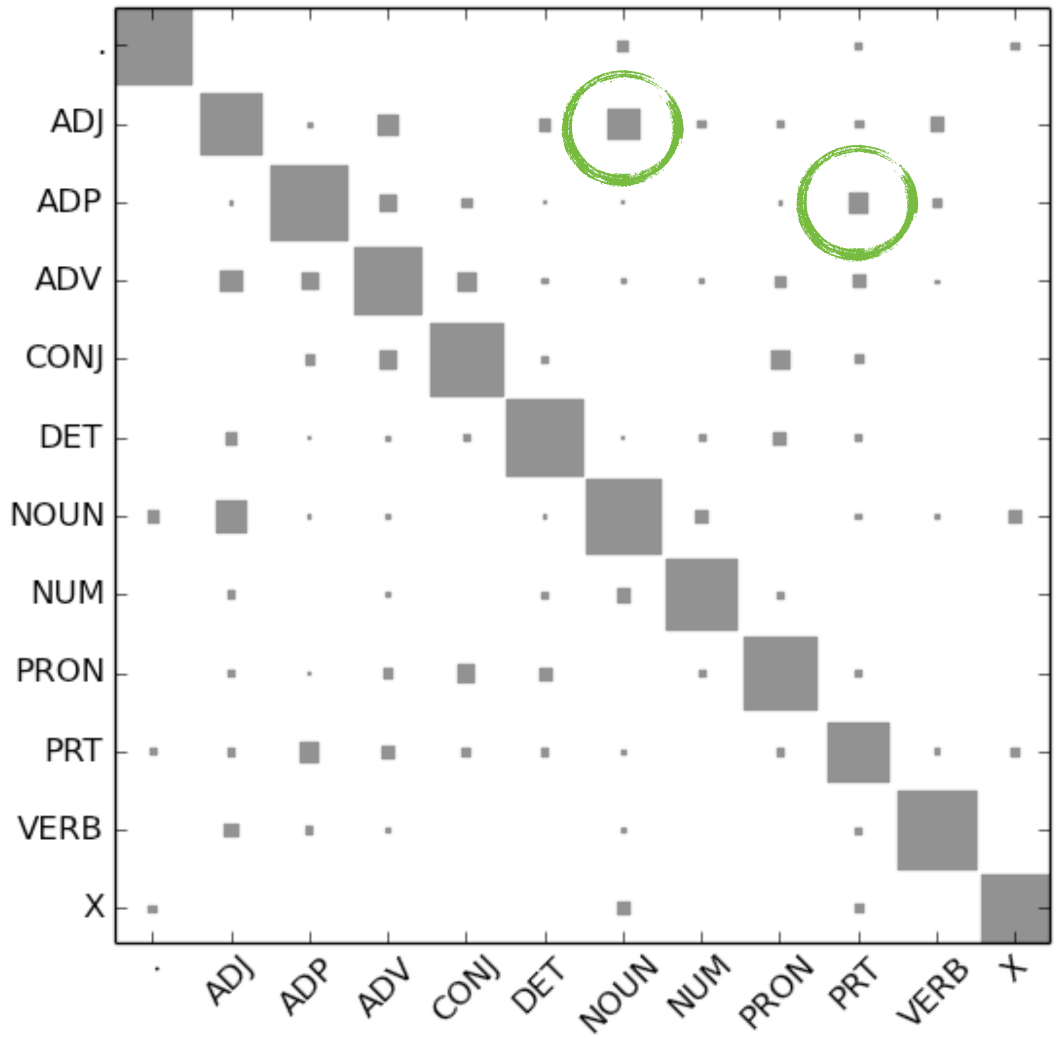


Wall Street Journal PTB-00

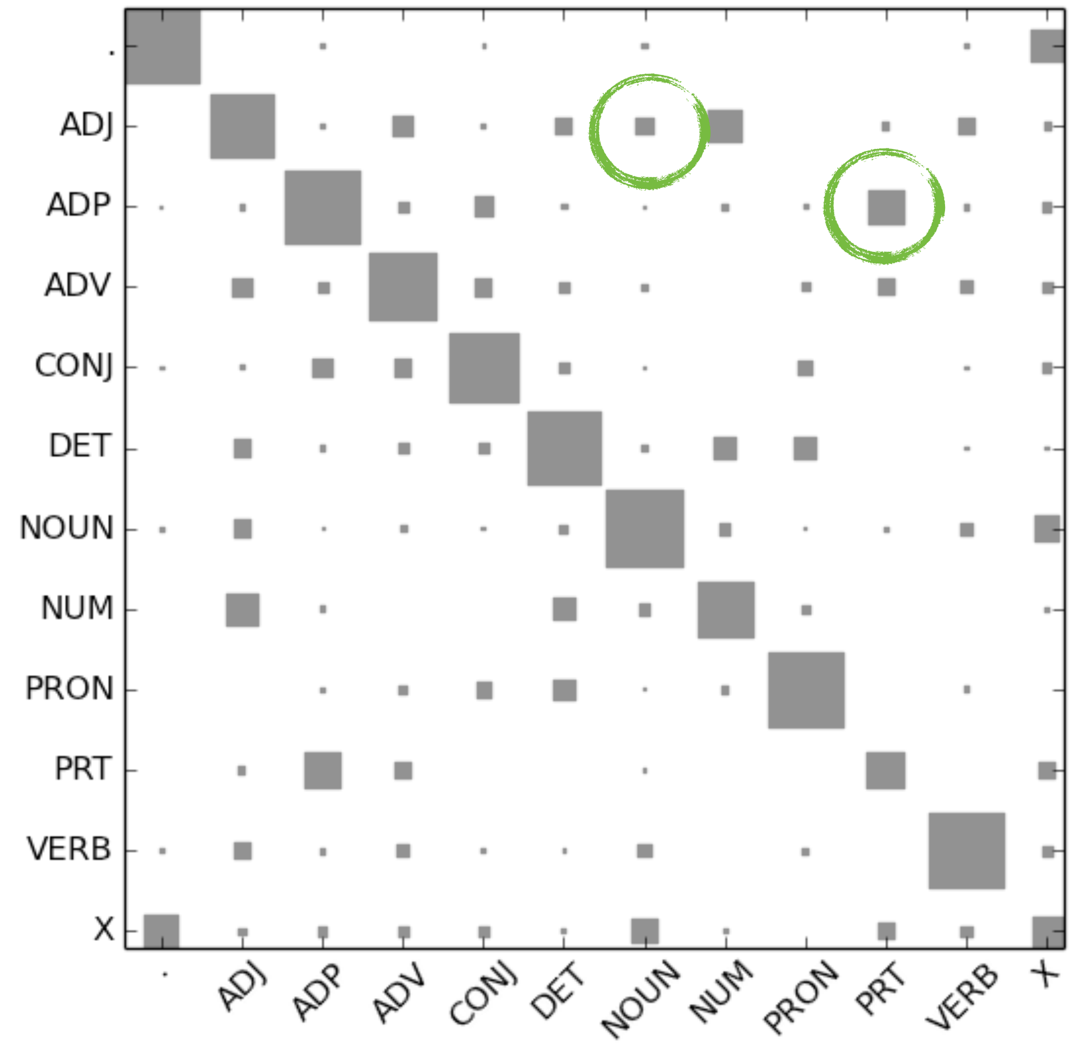


Twitter

(Plank et al., 2014)

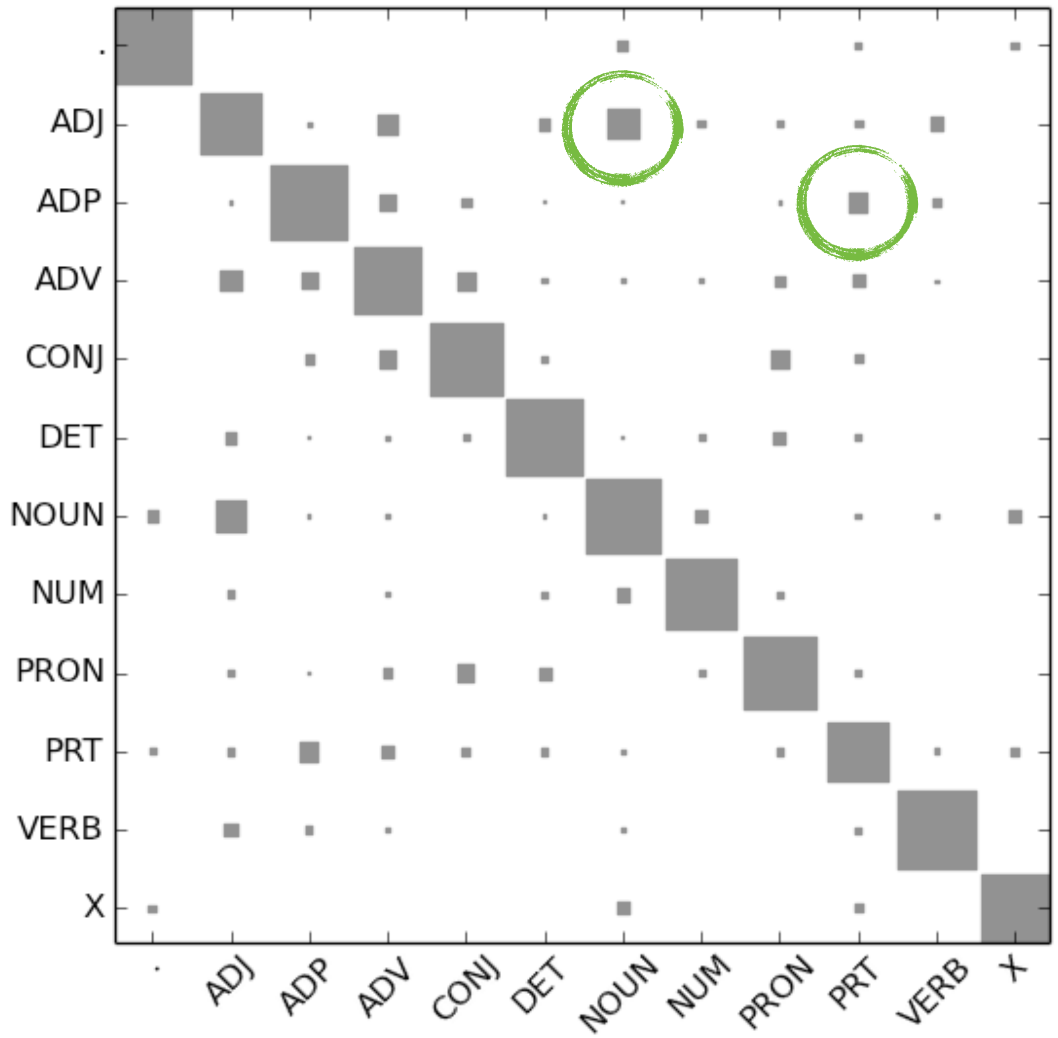


Wall Street Journal PTB-00

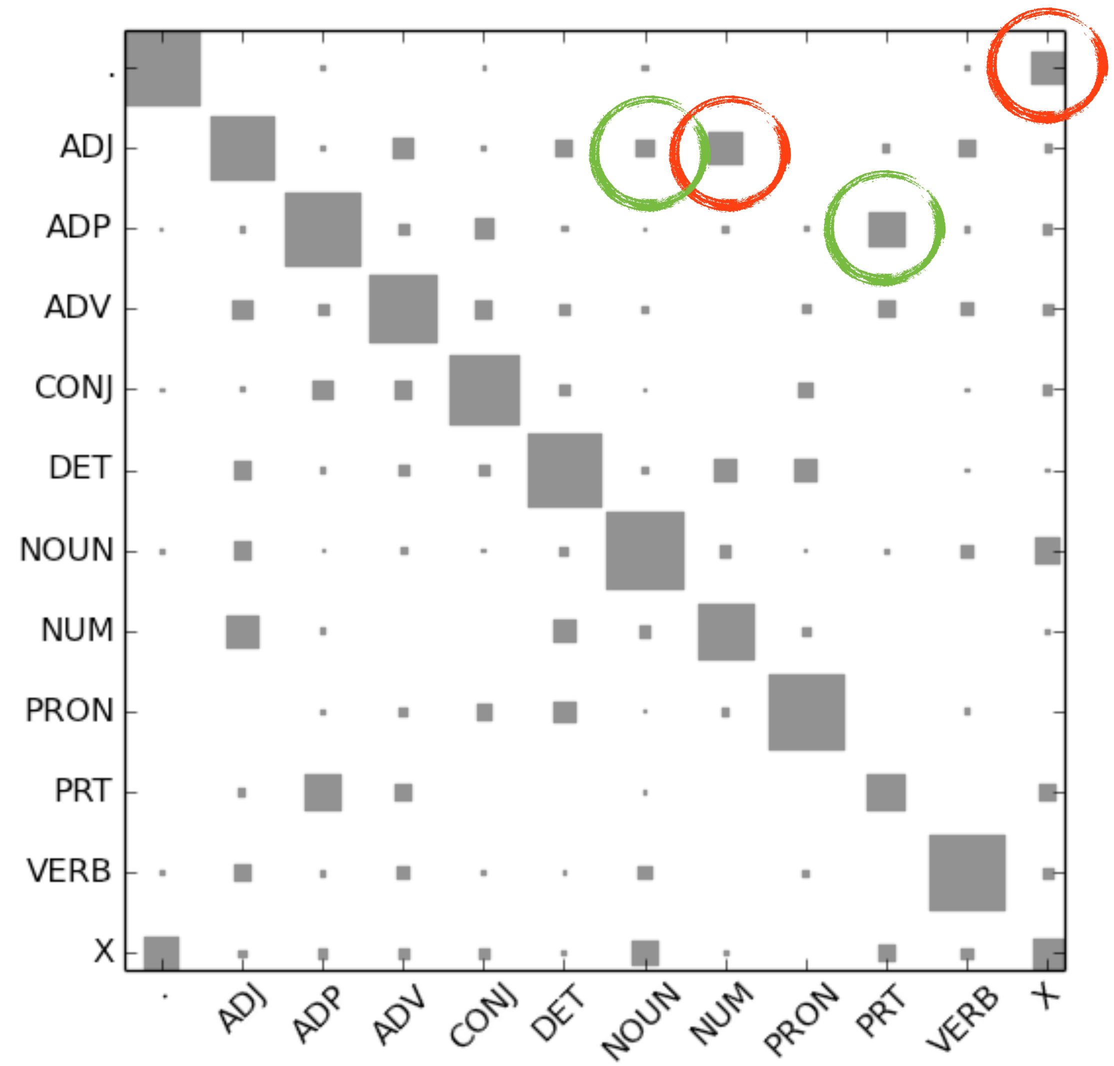


Twitter

(Plank et al., 2014)



Wall Street Journal PTB-00



Twitter

(Plank et al., 2014)

Are disagreements randomly distributed? No.

... and can we estimate disagreements from small samples? Yes!

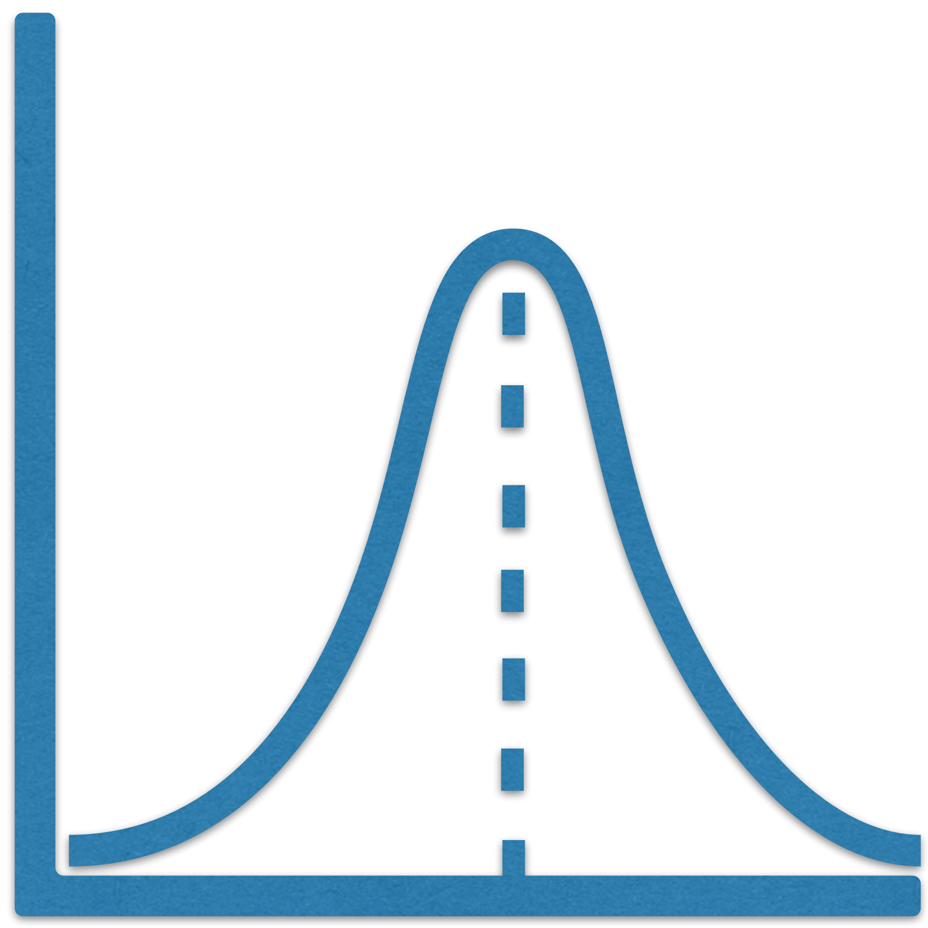
(Plank et al., 2014)

Are disagreement distributions unimodal?

… do they contain inherent disagreement signal?

(Pavlick & Kwiatkovski, 2019)

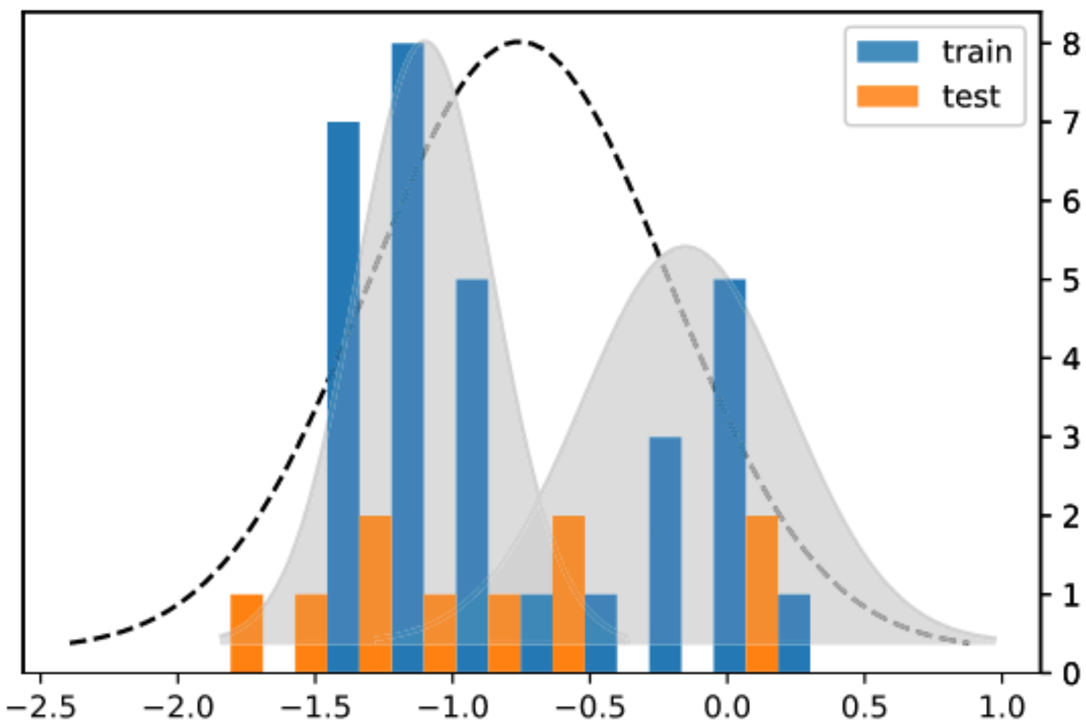
Is Unimodal (= Single Truth) Enough?



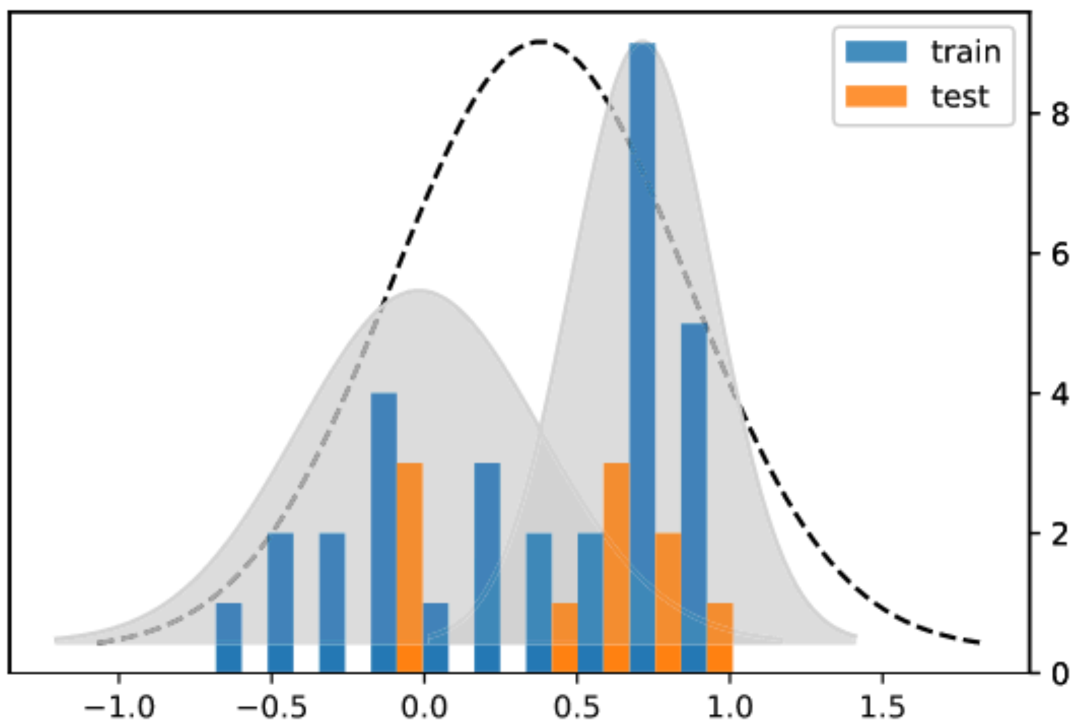
(Pavlick & Kwiatkovski, 2019)

Examples with bi-modal human judgement distributions

p: A homeless man being observed by a man in business attire.
h: Two men are sleeping in a hotel.



p: Paula swatted the fly.
h: The swatting happened in a forceful manner.

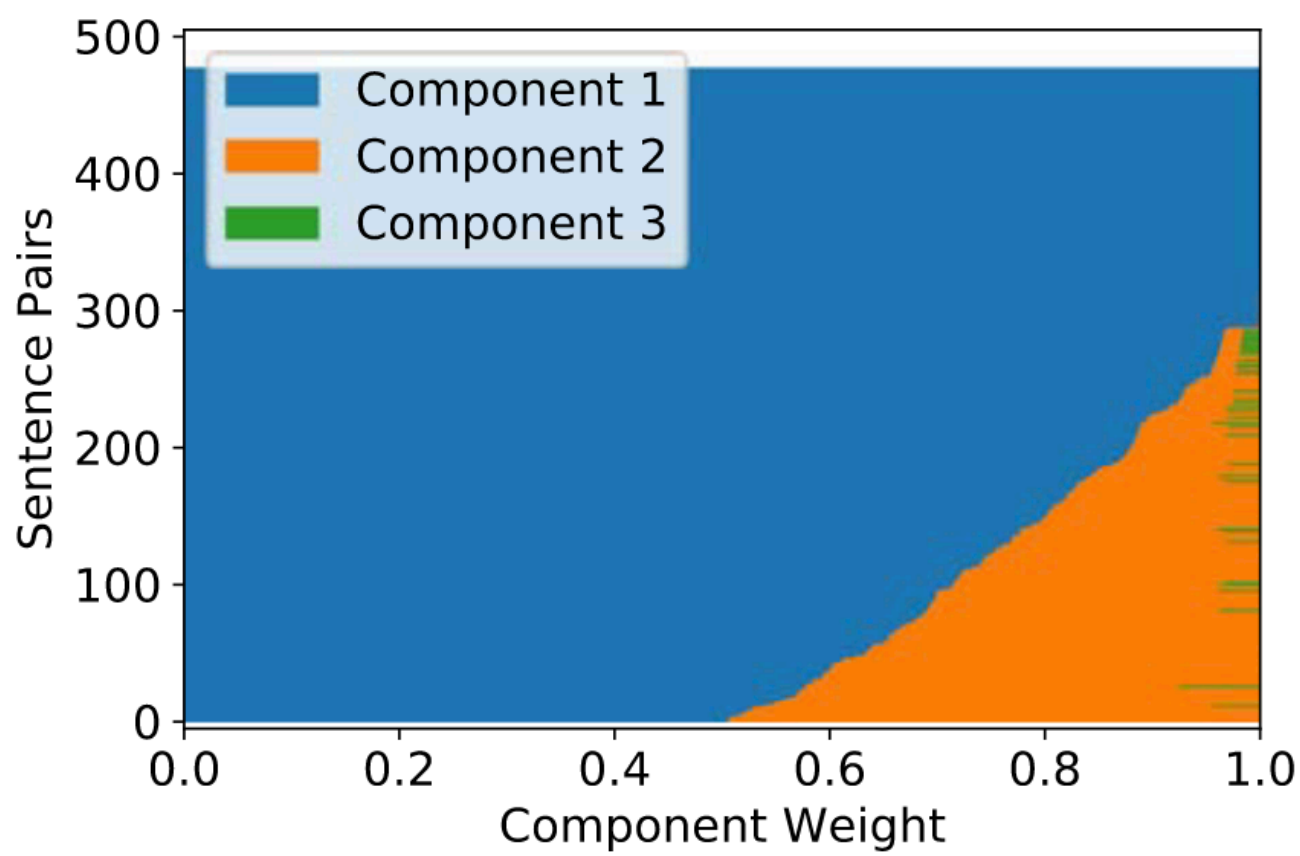


GMM with 1 *component* vs *k components*

(Pavlick & Kwiatkovski, 2019)

RTE Analysis

“For 20% of the sentence pairs, there is a non-trivial second component”



(Pavlick & Kwiatkovski, 2019)

Are disagreement distributions unimodal? No.

… do they contain inherent disagreement signal? Yes!

(Pavlick & Kwiatkovski, 2019)

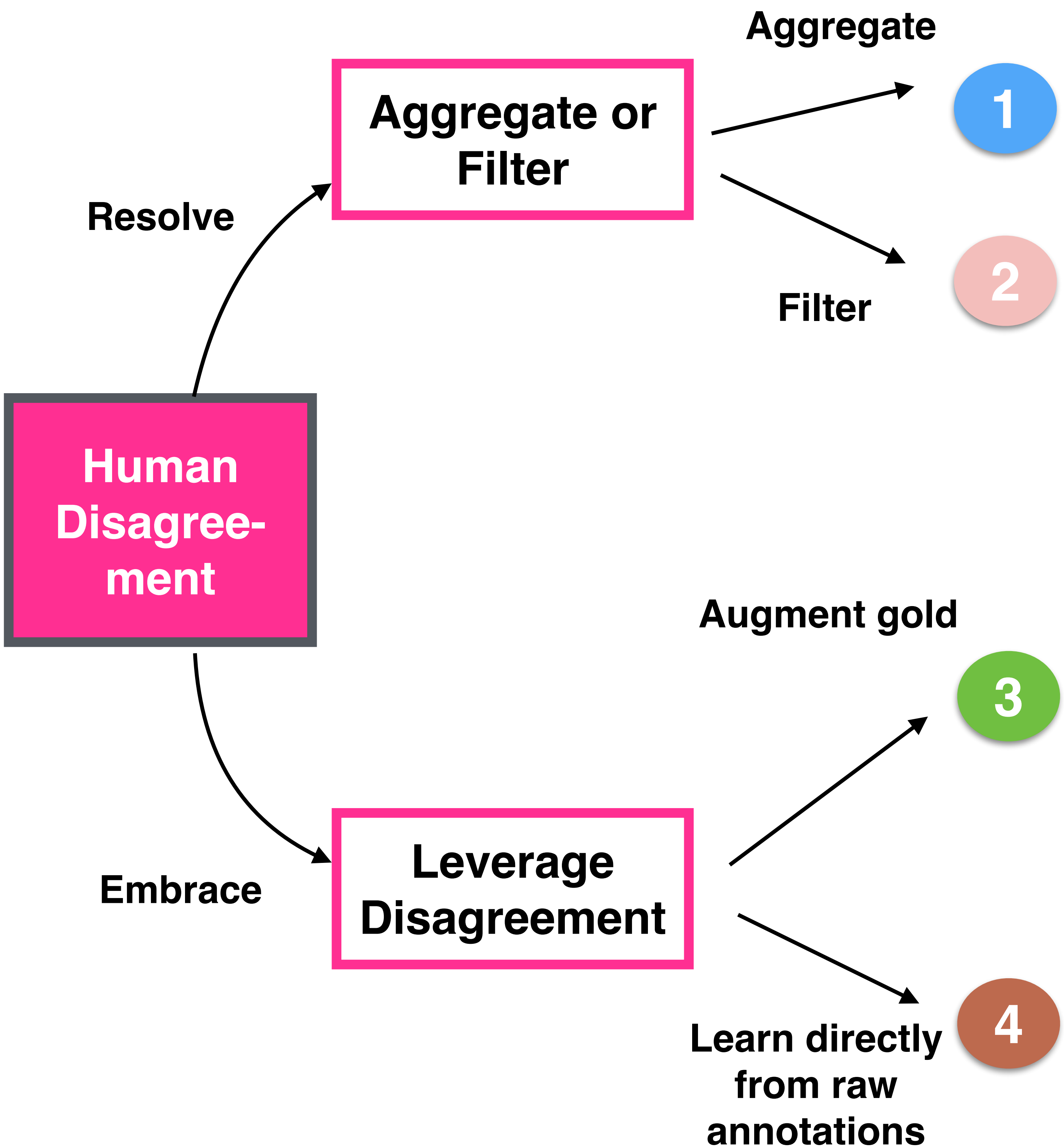
Roadmap

1. Data: Is disagreement random noise?
2. Modelling: How can we leverage disagreement?
3. Evaluation: How to evaluate in light of disagreement?

So what can we do?

Act II: Modelling

Learning with Disagreement



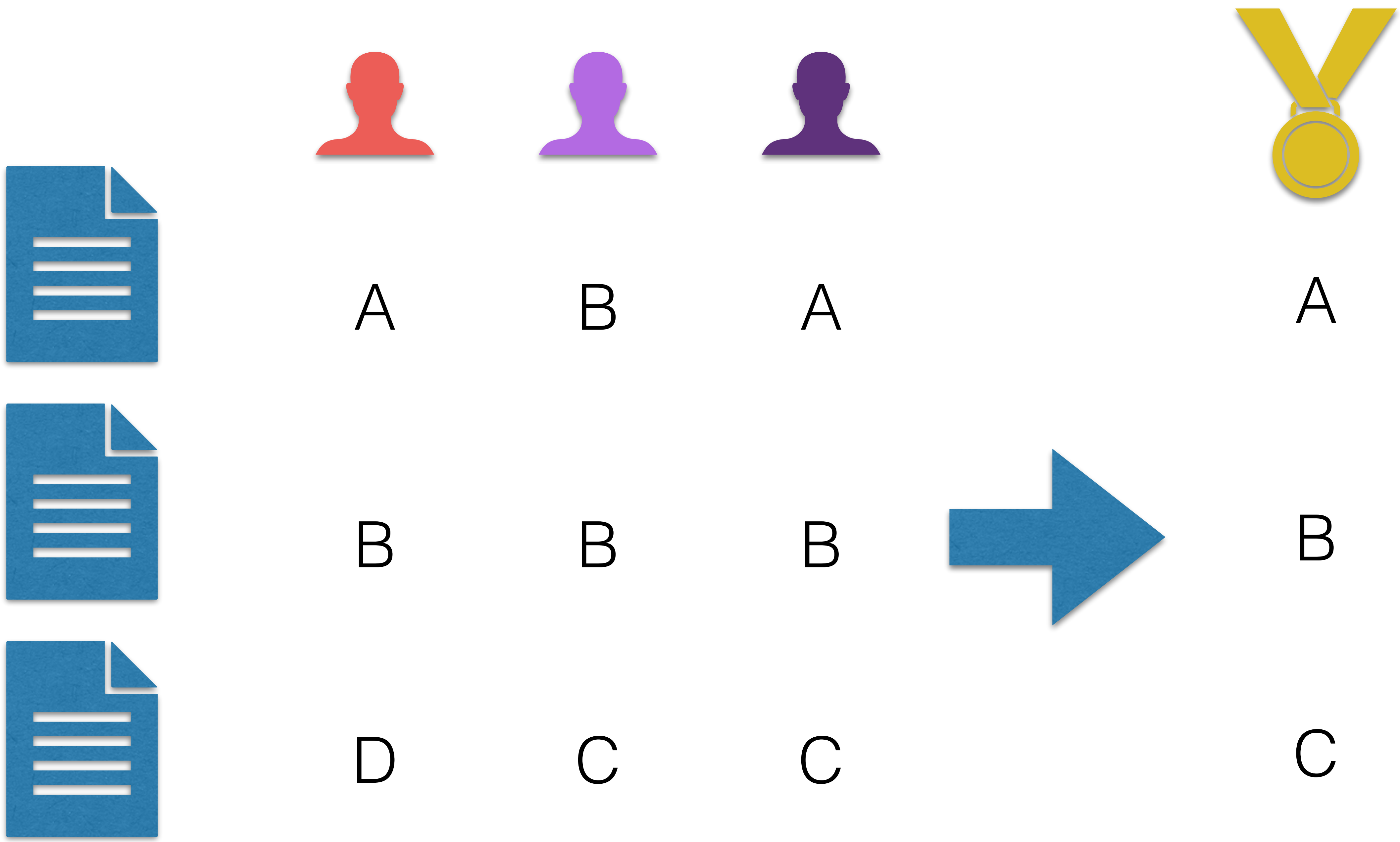
Aggregation methods
(e.g. Dawid & Skene, 1979;
Paun, Artstein, Poesio, forthcoming
Morgan & Claypool book)

Filter data
(e.g. hard filtering by Reidsema &
van den Akker, 2008; Beigman-Klebanov et al.)

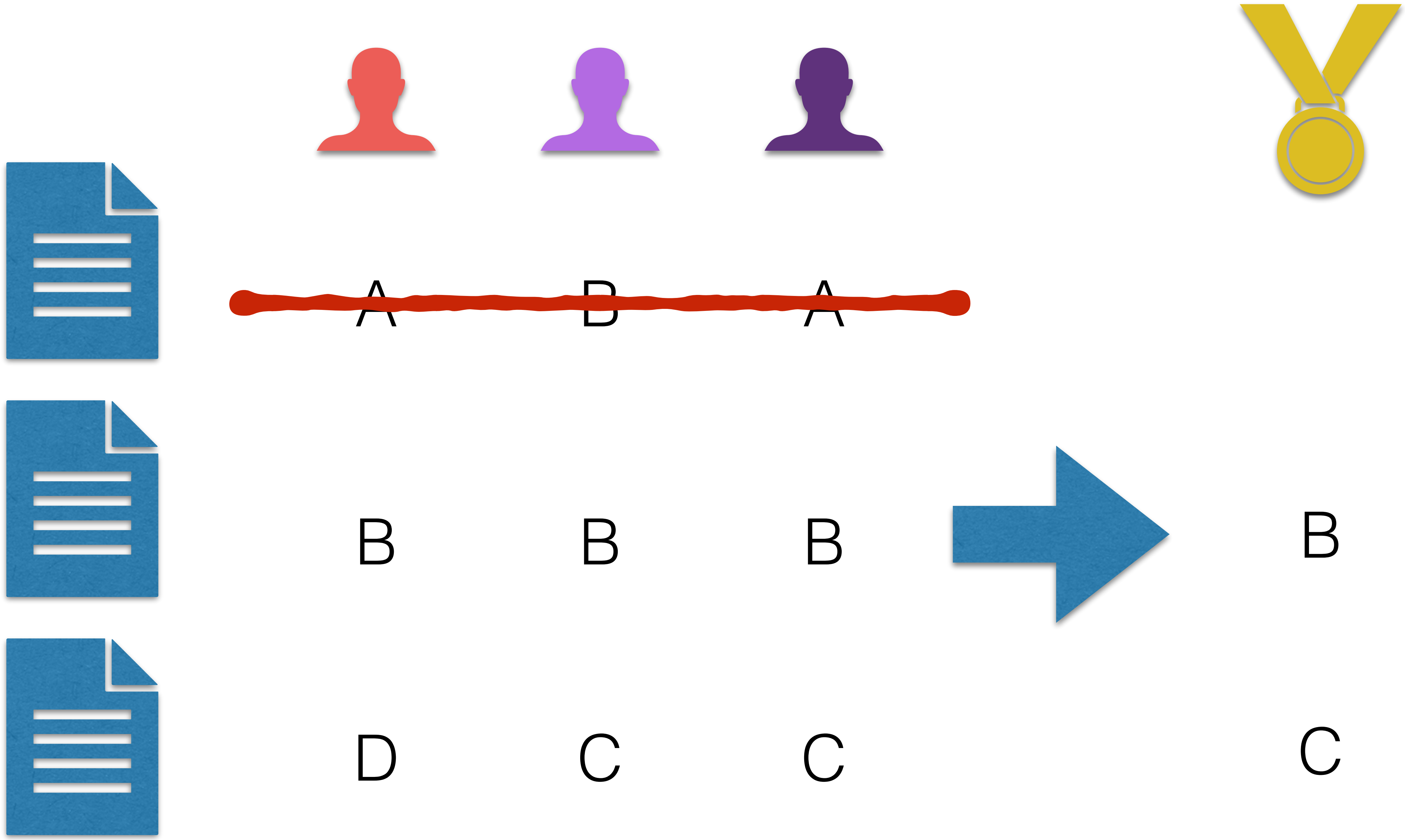
Weighting, Multi-task Soft Loss
(e.g. Plank et al., 2014; Fornaciari et al., 2021)

Soft-labels, CrowdTruth, DL from Crowd, Repeated Labeling
(e.g. Peterson et al., 2019, Uma et al. 2020;
Aroyo & Welty, 2014; Rodrigues & Peireia, 2017;
Sheng et al., 2008)

1 Aggregation

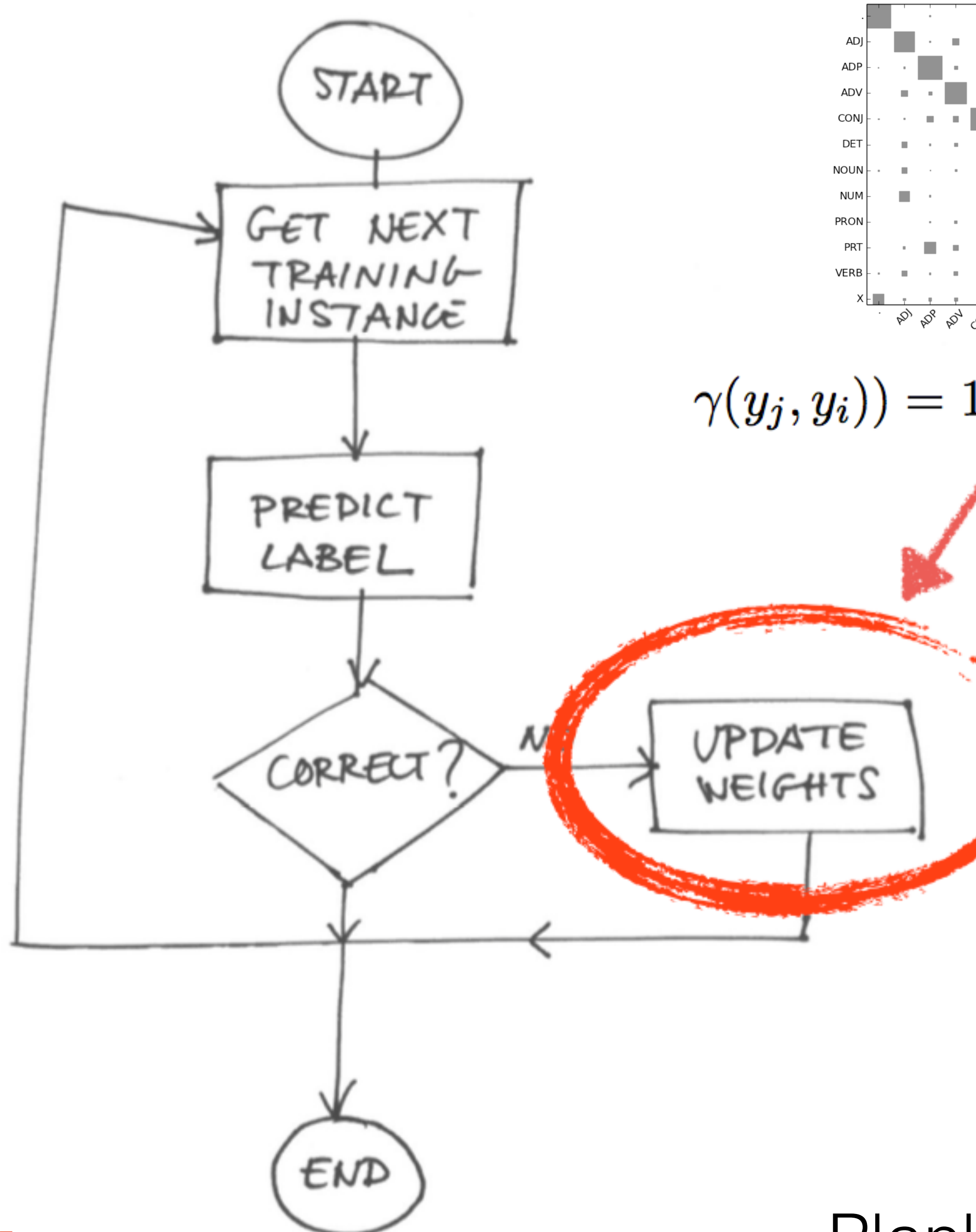


2 Filter

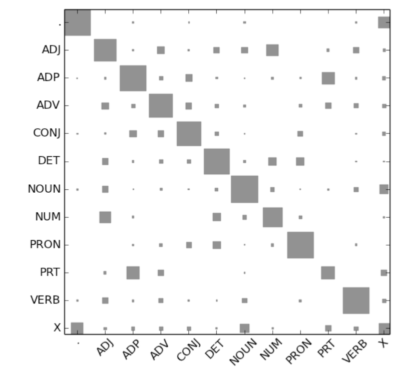


3 Augment gold with disagreement

Weighting by Disagreement



CM (confusion matrix)

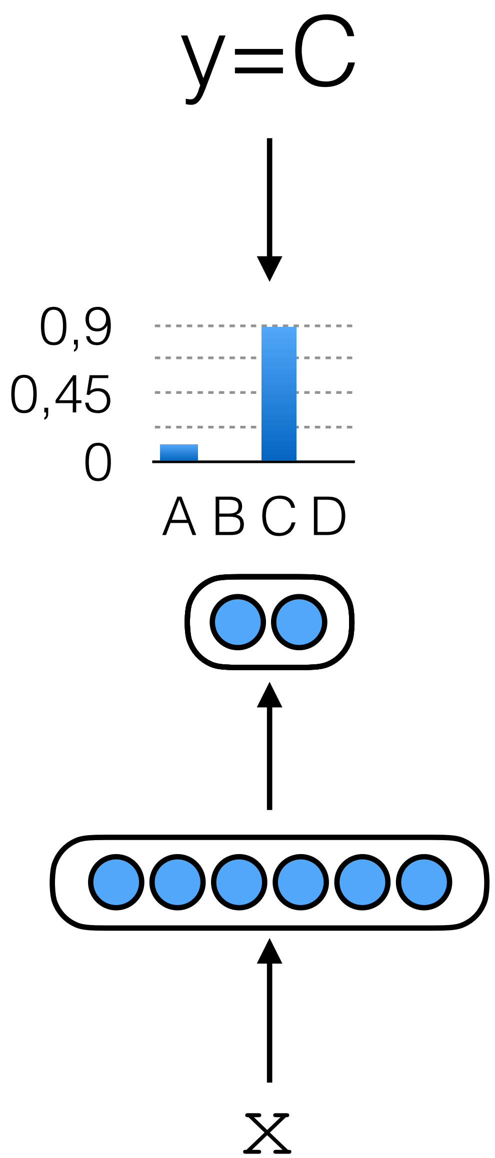


$$\gamma(y_j, y_i)) = 1 - P(\{A_1(X), A_2(X)\} = \{y_j, y_i\})$$

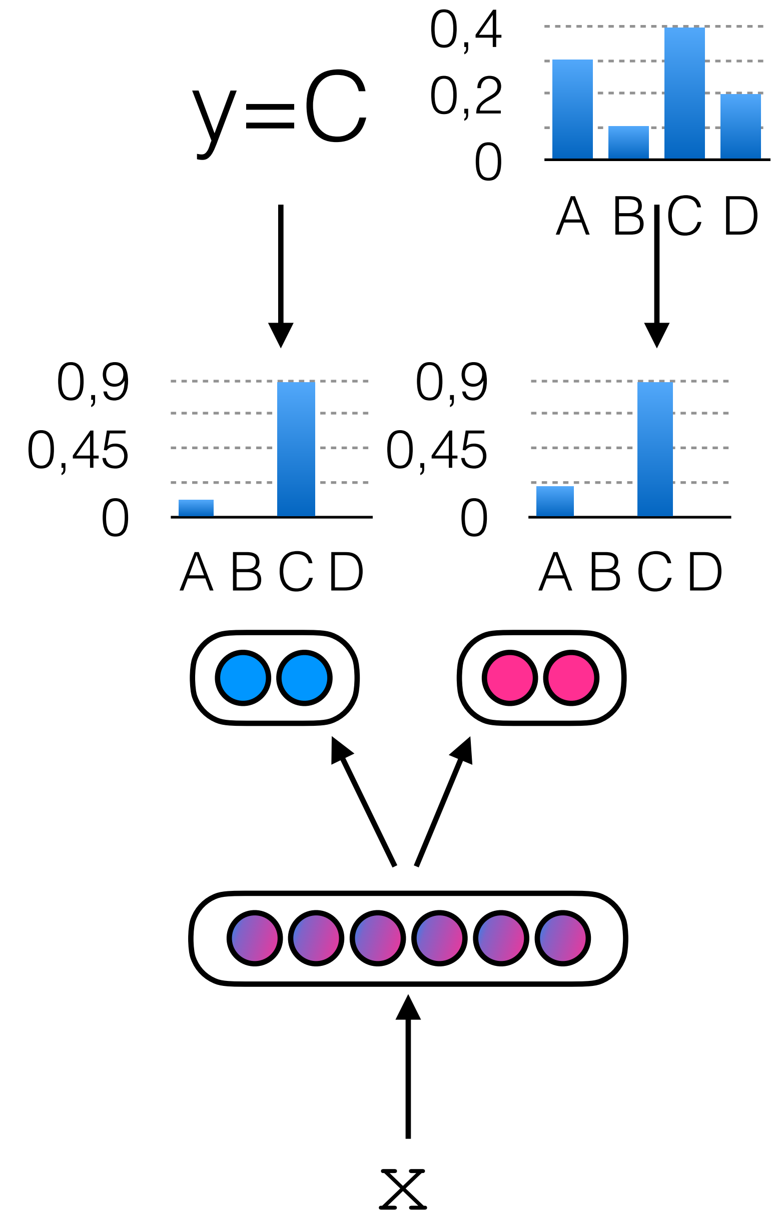
cost-sensitive learning

Plank, Hovy, Søgaard (2014)

Soft-labels via Multi-Task Learning



Gold label

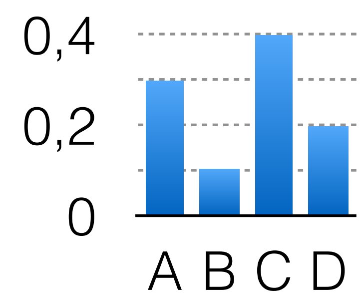


Gold label + Soft label

(Fornaciari, Uma, Paul, Plank, Hovy, Poesio 2021 NAACL)

Soft-labels

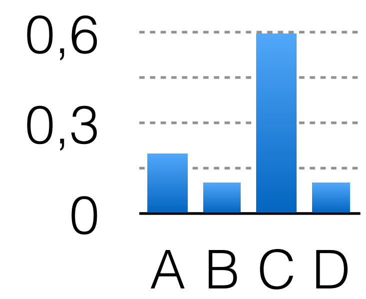
Annotator distribution P



Measure divergence

$$D_{KL}(P||Q) = \sum_i P(i)\, log_2 \left(\frac{P(i)}{Q(i)} \right)$$

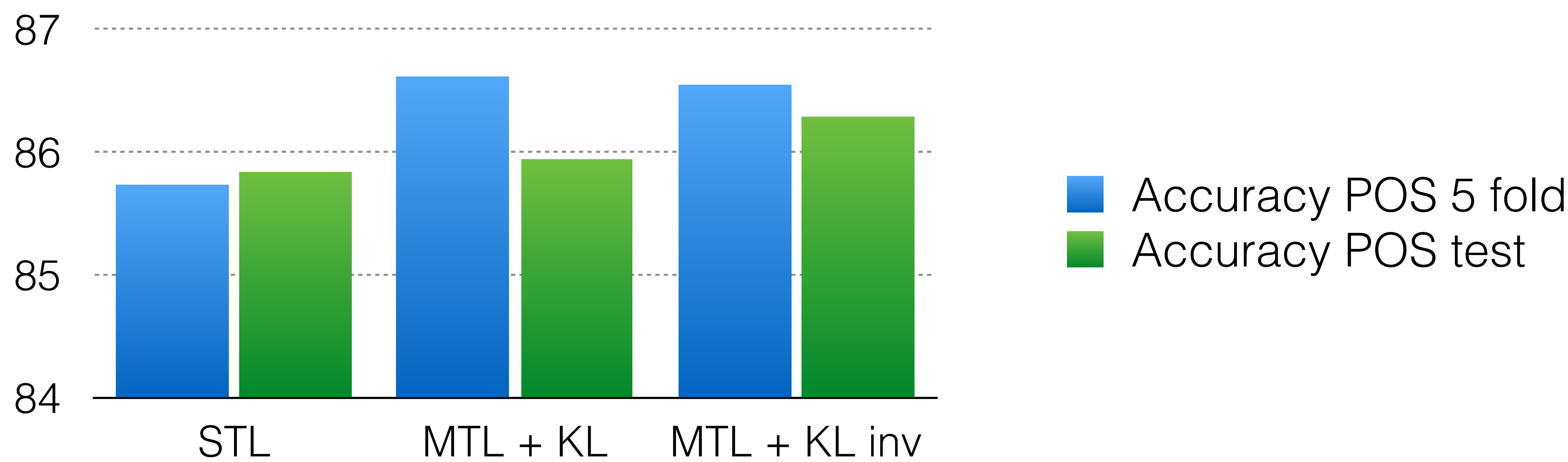
Predicted softmax Q

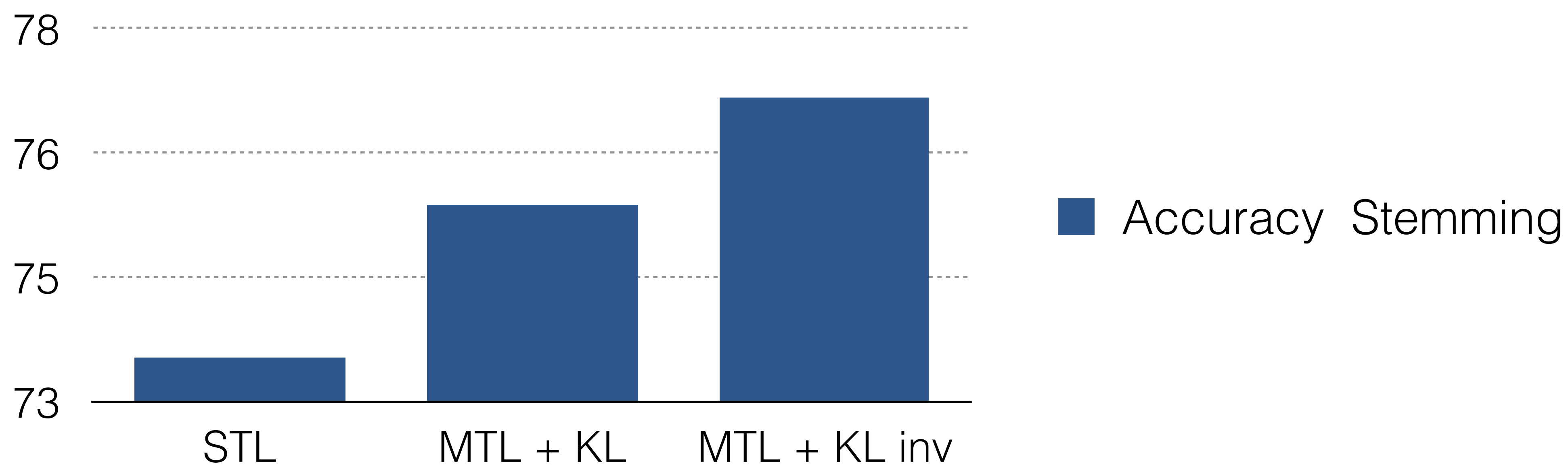


Experiments

- **Comparison:**
  - Single task learning
  - Multi-task learning (with gold or majority vote)
    - With soft loss
  - Two NLP tasks in this paper: POS and stemming

Results





$D_{KL}(P||Q)$ $D_{KL}(Q||P)$

4 Learn directly from raw annotations

E.g. Deep Learning from Crowd

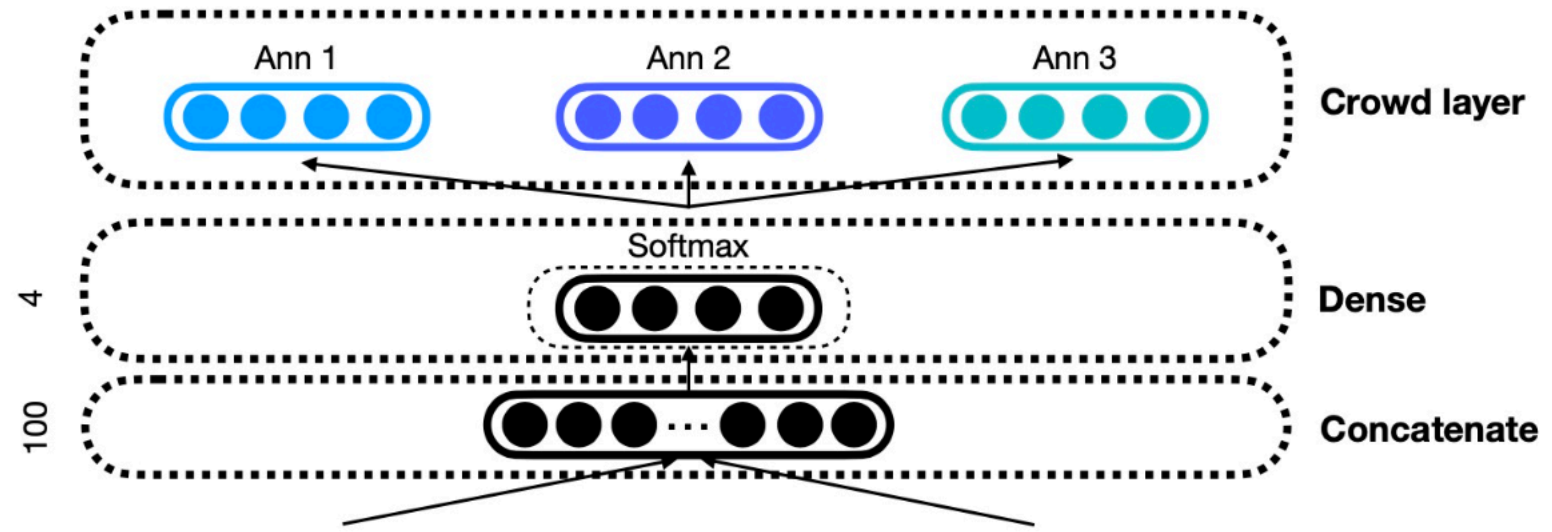


Figure 3: Illustration of deep learning from crowds proposed by Rodrigues and Pereira (2017).

Experiments: Understanding Indirect Questions

- **Dataset**: Friends-QIA dataset (Damgaard, Toborek, Eriksen & Plank, 2021) to appear in CODI @ EMNLP 2021; Fleiss 0.8833

Dataset	FRIENDS-QIA
All	5,930
Train	4,744
Dev	593
Test	593

All agree	75.02%
Two agree	23.39%
All disagree	1.59%

Table 3: Annotator agreement.

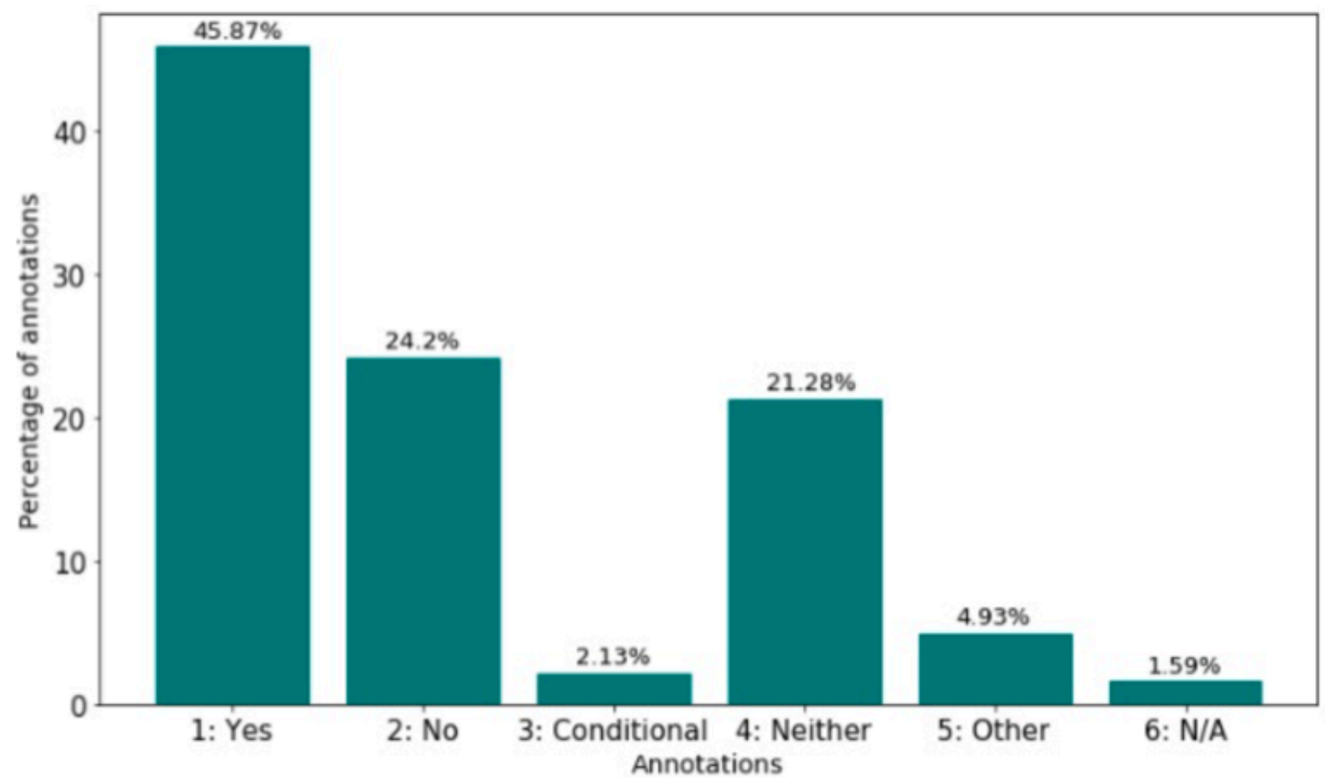
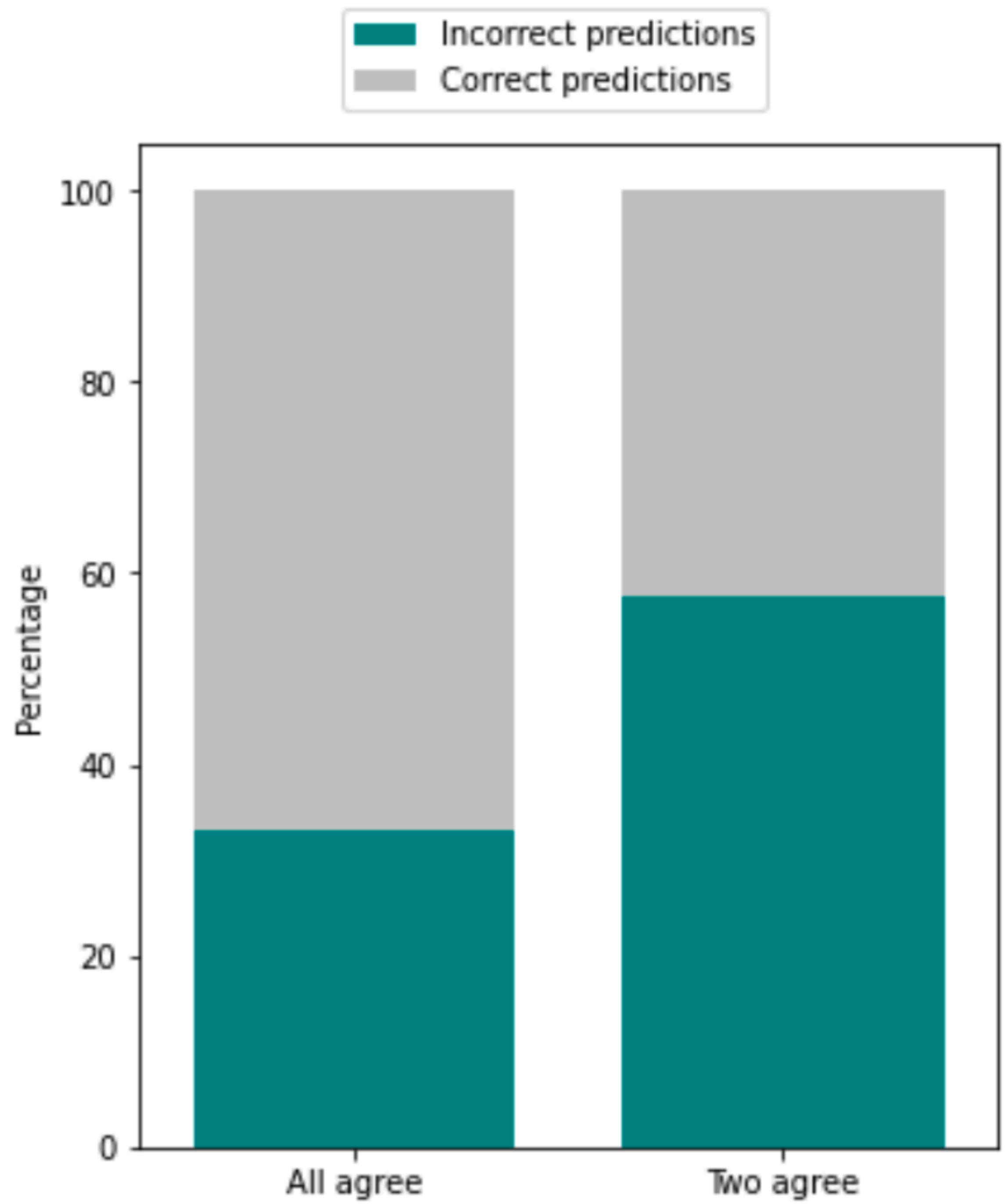


Figure 1: Gold label distribution.

Most incorrect predictions on instances humans did not agree on

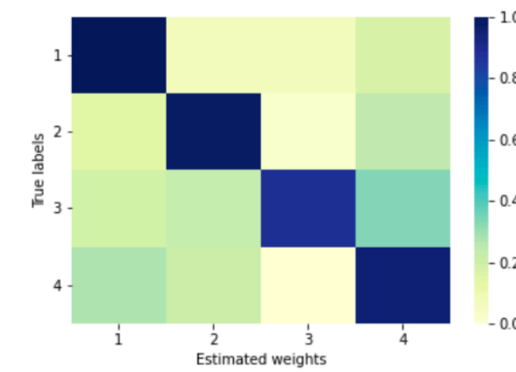


Correct and incorrect predictions of CNN with BERT vs. annotator agreement.

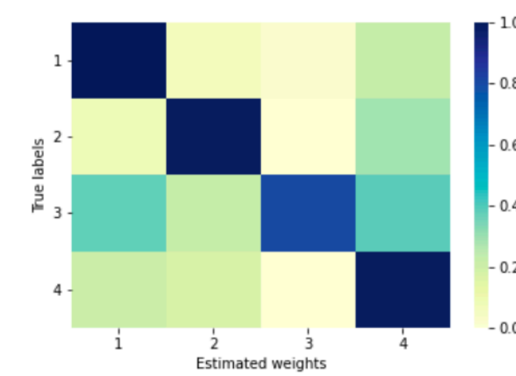
Deep Learning from Crowd

	Accuracy	F1-score
Majority baseline	49.07	16.46
Train on FRIENDS-QIA:		
CNN with BERT	**61.33**	45.65
CNN with BERT, multi-input	61.10	45.53
CNN with BERT + crowd layer	60.32	**47.89**
Train on FRIENDS-QIA + CIRCA:		
CNN with BERT	58.52	41.82

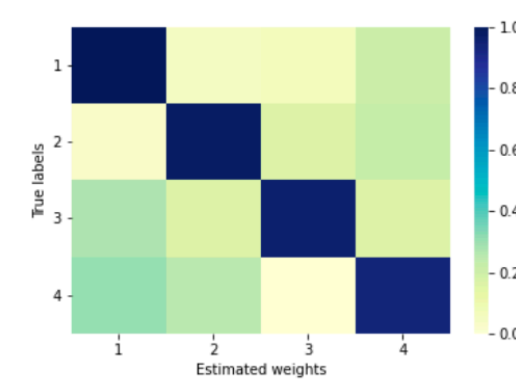
Table 7: Results on the FRIENDS-QIA test data.



(a) Annotator 1



(b) Annotator 2



(c) Annotator 3

More methods, overview and empirical evaluations:

JAIR survey by Uma et al., 2021:

Learning from Disagreement: A Survey

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio (2021 JAIR, forthcoming)

Roadmap

1. Data: Is disagreement random noise?
2. Modelling: How can we leverage disagreement?
3. Evaluation: How to evaluate in light of disagreement?

We Need to Talk about Disagreement in Evaluation

Work in collaboration with Alexandra Uma, Dirk Hovy, Massimo Poesio, Michael Fell, Silviu Paun, Tommaso Fornaciari, Valerio Basile (BPPF workshop@ACL 2021)

Evaluation in Interpretation Tasks

- A single correct answers ignores the **subjectivity** and **complexity** of many tasks

  ➡ Focus on “easy”, low-risk evaluation

- Many works on learning from disagreement compare against an evaluation set assumed to encode a **single ground truth**

Example from VQA 2.0

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

- Gold labels are often an **idealisation**, unreconcilable disagreement is abundant

Sources of disagreement

- **Stimulus characteristics** (ambiguity, task difficulty)
- **Individual differences** (incl. cultural and socio-demographics): for example in hate speech or sentiment
- **Context and attention** (Intra-coder disagreement; attention slips play a non-negligible role as well (Beigman Klebanov et al., 2008)

Similar position:

- Plank et al., (2014): Linguistically debatable or just plain wrong?
- Jamison & Gurevych (2015), Fornaciari et al., (2021): Noise or additional information?
- Aroyo & Welty (2015): Truth is a lie: Crowd Truth and the Seven Myths of human annotation
- Palomaki et al. (2018): a range of "acceptable variation"
- Uma et al. (2020), Basile (2020): Soft loss in NLP, evaluation

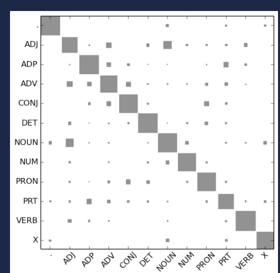
In contrast:

- Bowman & Dahl (2021): study and eliminate biases and artefacts in data
- Beigman Klebanov & Beigman (2009): evaluate on "easy" instances

Evaluation in Light of Disagreement

- Proposal: evaluate against hard *and* **soft labels**
- Soft evaluation sheds more light if uncertainty in models is similar to **human uncertainty** in labeling (**human collective**)
- Soft label evaluation e.g.:
  - Jensen-Shannon divergence (Uma et al., 2020; 2021; Nie et al., 2020); Uma et al. present further inf.-theoretic measures
  - Cross-entropy: Image classification (Peterson et al., 2019); in NLP (Pavlick & Kwiatkovski, 2019; Uma et al., 2020)
- Comparison of hard & soft evaluation in our upcoming survey

Take-home message



✓not all disagreement is noise

✓embrace it during learning

➡ Consider releasing raw annotations

✦ More work needed to understand forms of disagreement and embrace it in evaluation - see Uma et al. 2021 & Basile et al. 2021

IT UNIVERSITY OF COPENHAGEN

Questions? Thanks!

What to do about Human Disagreement in NLP?

@barbara_plank

bapl@itu.dk

nlpnorth.github.io

Thanks to the support by:

DANMARKS FRIE FORSKNINGSFOND

amazon

NVIDIA