

Disagreement in NLI

Jiang & de Marneffe (2022)

Fenghua Liu

L101 seminar · 12 November

Natural language inference

Given a premise P, assess a hypothesis H.

P: A child is running in the park.

Entailment (E): A child is outdoors.

Contradiction (C): No child is running.

Neutral (N): The child is happy.

One pair, several judgments

P: … they squeezed about 40000 people in there.

H: It does not hold many people.

What counts as “many”?

3 E, 0 N and 2 C votes.

UNIVERSITY OF CAMBRIDGE

Motivations

Why investigate disagreement?

NLI evaluation depends on annotation quality.

Majority labels hide plausible interpretations.

The causes of disagreement lack systematic analysis.

Why categorical models?

Vote distributions can reflect different kinds of uncertainty.

Applications such as fact-checking need categorical decisions.

UNIVERSITY OF CAMBRIDGE

Part I · Annotating reasons for disagreement

510 MNLI items

- 400 ChaosNLI items (100 votes each)
  one expert develops the taxonomy of reasons
  multiple reasons allowed per item
- 50 additional ChaosNLI items + 60 additional MNLI items (5 votes each)
  two experts annotate
- Agreement: MASI-based $\alpha = 0.69$

10 categories, 3 organising classes

Sentence meaning: lexical, implicature, presupposition, probabilistic enrichment, imperfection

Guidelines: coreference, temporal reference, interrogative hypothesis

Behaviour: accommodating added content, high overlap

UNIVERSITY OF CAMBRIDGE

What does a label distribution represent?

Category [4] “Probabilistic enrichment”

P: “It’s absurd …” Sir James nodded.

H: Sir James thinks it is absurd.

MNLI: [3,2,0] ChaosNLI: [63,35,2]

Does nodding indicate agreement?

Probabilistic enrichment gives different judgments.

Category [7] ”Temporal Reference”

P: Co-requesters cannot restrict timing after the release is issued.

H: They cannot restrict timing.

MNLI: [3,2,0] ChaosNLI: [90,8,2]

Evaluate H before or after release?

Fixing underspecified context gives different judgements.

UNIVERSITY OF CAMBRIDGE

Part II · Two categorical models

Four-way classification

RoBERTa → one output:

E / N / C / Complicated

Standard classification training

“Complicated” identifies items with low agreement.

Multilabel classification

independent prediction: sigmoid on three logits from RoBERTa; include a label when its score > 0.5

-> a set, e.g. {E, N}

Train with cross-entropy

The set identifies which labels compete.

UNIVERSITY OF CAMBRIDGE

Training targets and comparison

Gold labels constructed from votes

Four way:
>80% majority votes → E/N/C

<60% majority votes → Complicated

[60%, 80%] → excluded

Multilabel: include labels with ≥20% votes

5,482 items after augmentation

Train / development / test:

2,710 / 816 / 1,956

MixUp baseline: train on vote distributions using KL divergence.

Convert its probabilities ≥0.2 to sets.

Compare both on categorical targets.

UNIVERSITY OF CAMBRIDGE

Four-way classification

Test metric (%)	MixUp	4-way
Macro F1	65.34	**68.59**
Complicated F1	48.76	**62.32**

Multilabel classification

Test metric (%)	MixUp	Multi
Macro precision	**84.25**	81.19
Macro recall	76.81	**83.35**
Macro F1	80.29	**82.17**
2-label exact match	30.20	**46.16**

UNIVERSITY OF CAMBRIDGE

Result interpretation and error analysis

complicated label has lower F1 than other labels in the four-way model.

multilabel model is better than four-way because it is more *expressive*.

Takeaway: *heterogeneous* labels like “complicated” are hard to grasp for models and should be encoded alternatively.

Multilabel model recalls reasons of disagreement rather than literal >20% labels => better recall, worse precision

P: Oh, sorry, wrong church.

H: He or she entered the wrong church.

[82,17,1] => gold {E} 17% < 20% cut-off

Multilabel prediction: {E, N}

Reason category: probabilistic enrichment

Takeaway: a model trained on disagreement data can learn patterns of disagreement.

UNIVERSITY OF CAMBRIDGE

Critique · How principled is the taxonomy?

Intuitive but fluffy taxonomy of reasons

Their three major classes:

- sentence meaning uncertainty
- underspecifications in guidelines
- annotator behaviour

But uncertainty gives room for behavioural variation, which the guidelines fail to regulate…

“Expertise” ≠ reliable reasoning

Only 110 items were double-annotated, yet almost half of these exhibit disagreement.

MASI distance $\alpha = 0.69$.

Final analysis verifies usefulness of a reason, any reason, but does not verify usefulness of internal taxonomy of reasons.

UNIVERSITY OF CAMBRIDGE

Critique · Bad linguistic/behavioural analysis

Table 1 [6]

P: This was built 15 years earlier by Jahangir's wife, Nur Jahan, for her father, who served as Mughal Prime Minister.

H: Nur Jahan's husband Jahangir served as Mughal Prime Minister.

UNIVERSITY OF CAMBRIDGE

Critique · Bad linguistic/behavioural analysis

Table 1 [6]

P: This was built 15 years earlier by Jahangir's wife, Nur Jahan, for her father, who served as Mughal Prime Minister.

H: Nur Jahan's husband Jahangir served as Mughal Prime Minister.

[2,0,3] [24,45,31]

Their analysis: "the hypothesis is a **contradiction** if we assume that Mughal Prime Minister is the same person in both the premise and the hypothesis. However, it could be the case that Nur Jahan's father and husband both served as Mughal Prime Minister but in different terms, making it **Neutral**.

My analysis:

"served as Prime Minister" predicates a property of a person.

The father's having that property is semantically independent of the husband's having that property. => the correct label is **neutral**

why label **contradiction**?
"If Jahangir served as PM, the author would probably have mentioned it earlier, and then said 'for her father, who ALSO served as PM'." – a discourse-based implicature.

why label **entailment**? annotators simply didn't realise "father" vs "husband".

UNIVERSITY OF CAMBRIDGE

Critique · Two “unexplained” cases involving BASIC pragmatics

Example (6)

P: I’m confused.

H: Not all of it is very clear to me.

[92,5,3]

My analysis: scalar implicature

“not all …” can imply “some is clear”

+ lexical ambiguity
How much confusion characterises ’confused’?

Example (7)

P: You can’t do that in a skirt.

H: You can do anything in a skirt.

Proposed analysis: reference fixing for pronoun

[3,23,74]

P: ‘you’ => the addressee in the dialogue

H: if ‘you’ = annotator => N

if ‘you’ = generic ‘one’ / the addressee in the dialogue => C

UNIVERSITY OF CAMBRIDGE

Critique · Categorical ≠ decisive

The authors reject probabilistic classifier in favour of categorical classifier, but categoricity does not *entail* decisiveness:
a downstream claim verifier should not output “it is complicated” or “it is both entailed and neutral”.

spotting the problem √

solving the problem ×

UNIVERSITY OF CAMBRIDGE

Critique · Predicting human behaviour vs predicting inference validity

Population behaviour

question: Which labels do people select?

data: crowd-sourced judgements

e.g. case xxx: {E: 40%; N: 50%; C: 10%}

ideal output: {E: 40%; N: 50%; C: 10%} or {E, N}

application: behavioural research

Linguistic and task adequacy

question: which labels are correct/beneficial?

data: linguist-corroborated single judgements

e.g. case xxx: N

ideal output: N

application: downstream tasks like claim verifier, fact checker etc.

Other critiques

- **Weak link between the two parts**
  Reason annotations are used only in evaluation not in learning
- **Test distribution prediction as well for fairness**
  Baseline is trained on distribution and evaluated on categorical prediction.
  Their models are trained on categorical prediction and evaluated on categorical prediction.
- **Arbitrary yet consequential thresholds**
  High agreement: >80%; Low agreement: <60%; Gold label membership: ≥20%;
  Model label membership: score >0.5
  => Small vote changes can change the target.
  Robustness tests could have been conducted.

UNIVERSITY OF CAMBRIDGE

Assessment

Claimed contribution	Assessment
Taxonomy of reasons	× for a linguist
Reusable disagreement annotations	? based on questionable guidelines + there is disagreement
Categorical disagreement models	√ implemented and evaluated

Outcomes	Assessment
Better interpretation recall	√ for this task
More expressive outputs	? reasons for disagreement not given
Downstream usefulness	? quoted as motivation but evidently not thought through