Investigating Reasons for Disagreement in Natural Language Inference

Nan-Jiang Jiang
Department of Linguistics
The Ohio State University,
USA
jiang.1879@osu.edu

Marie-Catherine de Marneffe
Department of Linguistics / FNRS
The Ohio State University / UCLouvain
USA / Belgium
demarneffe.1@osu.edu

Abstract

We investigate how disagreement in natural language inference (NLI) annotation arises. We developed a taxonomy of disagreement sources with 10 categories spanning 3 high-level classes. We found that some disagreements are due to uncertainty in the sentence meaning, others to annotator biases and task artifacts, leading to different interpretations of the label distribution. We explore two modeling approaches for detecting items with potential disagreement: a 4-way classification with a ''Complicated'' label in addition to the three standard NLI labels, and a multilabel classification approach. We found that the multilabel classification is more expressive and gives better recall of the possible interpretations in the data.

1 Introduction

Natural language inference (NLI) is the task of identifying whether a hypothesis sentence is true, false, or undetermined, given a premise. It is considered one of the most fundamental aspects of competent language understanding. In natural language processing, the NLI task is widely used to evaluate models' semantic representations (i.a., Wang et al., 2019a), and to facilitate downstream tasks, for example, in natural language generation (NLG).

Large NLI datasets have been built by collecting inference judgments for premise-hypothesis pairs and aggregating the judgments by simple methods such as majority voting. However, it has been pointed out that NLI items do not all have a single ground truth and can exhibit systematic disagreement (i.a., Pavlick and Kwiatkowski, 2019; Nie et al., 2020). This questions the assumption of having a single ground truth for each item and the validity of measuring models' ability to produce such ground truth. For instance, in (1) from the MNLI dataset (Williams et al., 2018), 3 out of 5 annotators labeled the item as ''Entailment'' (the hypothesis is inferred from the premise), 0 labeled it as ''Neutral'' (the hypothesis cannot be inferred from the premise), and 2 as ''Contradiction'' (the hypothesis contradicts the premise).

(1) **P:** the only problem is it's not large enough it only holds about i think they squeezed when Ryan struck out his five thousandth player they they squeezed about forty thousand people in there.

H: It doesn't hold many people. [E,N,C]: [3,0,2]

People have indeed different judgments on which number is required to count as *holding many people*. The premise and hypothesis do not resolve explicitly what is being talked about, possibly a stadium. Does 40,000 count as *many* for a stadium seating capacity? The premise states that *it's not large enough* and uses the term *squeezing*, leading some annotators to see the hypothesis *it doesn't hold many people* as being inferred from the premise. On the other hand, 40,000 people in a specific location is a large number, and some annotators therefore judge the hypothesis as contradictory to the premise. Such disagreement is not captured when taking only one of the three standard NLI labels as ground truth. Recent work (Zhang et al., 2021; Zhou et al., 2021) has thus explored approaches for building NLI models that predict the entire annotation distribution, instead of the majority vote category, in an attempt to move away from assuming a single ground truth per item. However, little is understood about where the disagreement stems from, and whether modeling the distribution is the best way to handle disagreement in annotation.

To investigate these questions, we created a taxonomy of different types of disagreement consisting of 10 categories, falling into 3 high-level classes based on the ''Triangle of Reference'' by

Transactions of the Association for Computational Linguistics, vol. 10, pp. 1357–1374, 2022. https://doi.org/10.1162/tacl_a_00523
Action Editor: Anette Frank. Submission batch: 3/2022; Revision batch: 7/2022; Published 12/2022.

Aroyo and Welty (2015). We manually annotated a subset of MNLI with the 10 categories. Our categorization shows that items leading to disagreement in annotation are highly heterogeneous. Moreover, the interpretation of the NLI label distribution differs across items. We thus explored alternative approaches for modeling disagreement items: A 4-way classification approach with an additional label (on top of the three NLI labels) capturing disagreement items, and a multilabel classification approach of predicting one or more of the three NLI labels. We found that the two models behave somewhat differently, with the multilabel model offering more interpretable outputs, and thus being more expressive. Our findings deepen our understanding of disagreement in a widely used NLI benchmark and contribute to the growing literature on disagreement in annotation. We hope they highlight directions to reduce disagreement when collecting annotations and to design models to handle the disagreement that persists. The annotations, the guidelines, and the code are available at `https://github.com/njjiang/NLI_disagreement_taxonomy`.

2 Related Work

Focusing on disagreement in annotation is not new: Aroyo and Welty (2015) argued for embracing annotation disagreement, viewing it as signal, and not as noise. Even for tasks with supposedly a unique correct answer, such as part-of-speech tagging, there are items for which the right analysis is debatable (Plank et al., 2014b): Is *social* in *social media* a noun or an adjective? Plank et al. (2014a) showed that incorporating such disagreement signals into the loss functions of part-of-speech taggers improves performance. Previous work noted that disagreement in annotation exists in many semantic tasks: Anaphora resolution (Poesio and Artstein, 2005; Versley, 2008; Poesio et al., 2019), coreference (Recasens et al., 2011), sentiment analysis (Kenyon-Dean et al., 2018), word sense disambiguation (Erk and McCarthy, 2009; Passonneau et al., 2012), among others.

Sources for Disagreement Aroyo and Welty (2015) introduced the ''Triangle of Reference'' framework to conceptualize the annotation process and explain annotation disagreement. Annotation differences can stem from the sentences to be annotated, the labels, or the annotators. Indeed, annotators, who interpret the sentences, produce labels in a way that is defined by the annotation guidelines. Underspecification in each of these three components can result in disagreement in the annotations. Disagreement can arise from (1) uncertainty in the sentence meaning, (2) underspecification of the guidelines, or (3) annotator behavior. We use the Triangle of Reference to organize our taxonomy.

Disagreement in NLI de Marneffe et al. (2012) and Uma et al. (2021) showed that disagreement was systematic in the older NLI datasets. Pavlick and Kwiatkowski (2019) showed that real-valued NLI annotations are better modeled as coming from a mixture of Gaussians as opposed to a single Gaussian distribution. Nie et al. (2020) collected categorical NLI annotations and found disagreement to be widespread, corroborating Pavlick and Kwiatkowski's (2019) findings. Kalouli et al. (2019) found that items involving entity/event coreference and ''loose definitions'' of inference (e.g., whether *a hill covered by grass* is the same as *the side of a mountain*) have lower inter-annotator agreement. However, there is not yet a systematic investigation of how disagreement in NLI arises.

Taxonomy in NLI There is a rich body of work on the taxonomy of reasoning types in NLI, identifying the kinds of inferences exhibited in NLI datasets (i.a., Sammons et al., 2010; LoBue and Yates, 2011; Williams et al., 2022). Our work differs in that we focus on the phenomena that lead to annotation disagreement, which are not necessarily reasoning types (e.g., our category Interrogative Hypothesis, [8] in Table 1). Since we focus on disagreement, we do not categorize different ways of arriving at the same NLI label (e.g., different kinds of high agreement contradiction, as in de Marneffe et al., 2008).

Approaches to Model Disagreement Pavlick and Kwiatkowski (2019) argued that NLI disagreement information should be propagated downstream. Current neural models should thus be evaluated against the full label distribution. Methods for approximating the full distribution have recently been developed for many tasks, using techniques for calibration and learning with soft-labels (i.a., Lalor et al., 2017; Zhang et al.,

		Premise	Hypothesis	MNLI [E,N,C]	ChaosNLI [E,N,C]
Uncertainty in Sentence Meaning					
[1]	Lexical	Technological advances generally come in waves that crest and eventually subside.	Advances in electronics come in waves.	[3,1,1]	[82,17,1]
[2]	Implicature	Today it is possible to buy cheap papyrus printed with gaudy Egyptian scenes in almost every souvenir shop in the country, but some of the most authentic are sold at The Pharaonic Village in Cairo where the papyrus is grown, processed, and hand-painted on site.	The Pharaonic Village in Cairo is the only place where one can buy authentic papyrus.	[0,2,3]	[2,39,41]
[3]	Presupposition	What changed?	Nothing changed.	[0,2,3]	[4,76,20]
[4]	Probabilistic Enrichment	It's absurd but I can't help it. Sir James nodded again.	Sir James thinks it's absurd.	[3,2,0]	[63,35,2]
[5]	Imperfection	profit rather	Our profit has not been good.	[0,3,2]	[3,90,7]
Underspecification in Guidelines					
[6]	Coreference	This was built 15 years earlier by Jahangir's wife, Nur Jahan, for her father, who served as Mughal Prime Minister.	Nur Jahan's husband Jahangir served as Mughal Prime Minister.	[2,0,3]	[24,45,31]
[7]	Temporal Reference	However, co-requesters cannot approve additional co-requesters or restrict the timing of the release of the product after it is issued.	They cannot restrict timing of the release of the product.	[3,2,0]	[90,8,2]
[8]	Interrogative Hypothesis	How did you get it?" A chair was overturned.	"How did you get your hands on this object?"	[3,2,0]	[45,52,3]
Annotator Behavior					
[9]	Accommodating Minimally Added Content	Indeed, 58 percent of Columbia/HCA's beds lie empty, compared with 35 percent of nonprofit beds.	58% of Columbia/HCA's beds are empty, said the report.	[3,2,0]	[97,3,0]
[10]	High Overlap	Yet, in the mouths of the white townsfolk of Salisbury, N.C., it sounds convincing.	White townsfolk in Salisbury, N.C. think it sounds convincing.	[3,2,0]	[68,27,5]

Table 1: Categories of potential sources of disagreement, with examples. The last two columns give the number of annotations for each NLI label ''Entailment'' (E), ''Neutral'' (N), and ''Contradiction'' (C), in MNLI and ChaosNLI.

2021; Fornaciari et al., 2021; Zhou et al., 2021; Uma et al., 2022).

However, simply because distributions are the most straightforward form of disagreement information does not mean that they are the optimal representation for intrinsic evaluation or in downstream tasks. Calibration techniques are successful at post-editing the classifier's softmax distribution (Guo et al., 2017), but they convey spurious uncertainty for items that do not exhibit disagreement (Zhang et al., 2021).

Categorical decisions tend to be more interpretable and are necessary in downstream tasks. For example, NLI models are often used for automatic fact-checking (Thorne et al., 2018; Luken et al., 2018), where the categorical decision of whether a statement is disinformation determines whether it needs to be censored. Therefore, we explore here different approaches for providing categorical information for disagreement.

For sentiment analysis, Kenyon-Dean et al. (2018) used a classification approach with an additional ''Complicated'' class to capture items with disagreement. Kenyon-Dean et al. (2018) had little success predicting that class with LSTM-based models (0.16 F1 for Complicated), because it is heterogeneous and there is likely few learning signals indicating complicatedness. Zhang and de Marneffe (2021) approached the NLI 4-way classification problem using the architecture of Artificial Annotator, an ensemble of multiple BERT models with different biases. They experimented on the NLI version of the CommitmentBank (de Marneffe et al., 2019), and showed some success, obtaining 61.93% F1 on the fourth class ''Disagreement'' using a vanilla-BERT baseline (standard fine-tuning BERT), and 66.5% F1 on the ''Disagreement'' class using the Artificial Annotator architecture. Here, we further test the 4-way classification approach for NLI.

In addition to its heterogeneity, a ''Complicated'' or ''Disagreement'' class is not easily interpretable. We not only need to know whether there is disagreement, but also in what way: Which labels do the annotators disagree over. We therefore also take a multilabel classification approach (i.a. Passonneau et al., 2012; Oh et al., 2019; Ferracane et al., 2021), predicting one or more of the three NLI labels.

There is another line of research aiming to model the judgments of individual annotators, as opposed to the aggregated annotations representing the judgments of the population (Gordon et al., 2021; Davani et al., 2022). However, these approaches require the annotators' identities for each annotation, which are often not released with the data.

3 Disagreement Taxonomy

To investigate where disagreement stems from, we conduct a qualitative analysis of parts of the

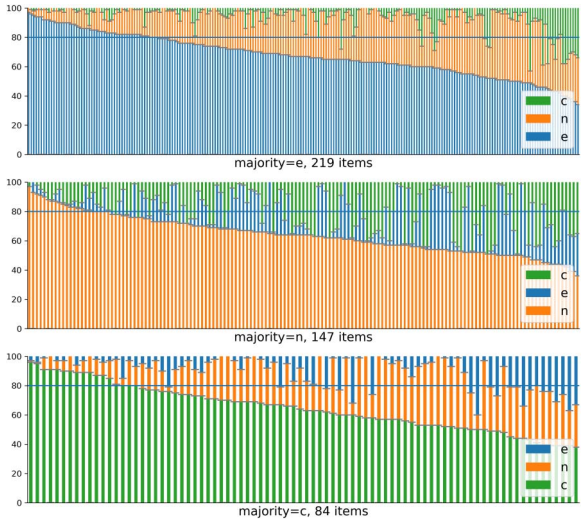

Figure 1: ChaosNLI annotations of the 450 items we sampled. Each column of stacked bars represents an item's annotations—the number of votes for each label with top-down ordering of the labels. The horizontal lines indicate 80 votes.

MNLI dataset (Williams et al., 2018). We chose MNLI because it is diverse in genre and inference types, compared with datasets based on image captions that only describe visual scenes (e.g., SICK [Marelli et al., 2014], SNLI [Bowman et al., 2015]).

3.1 Data to Analyze

The original MNLI dev sets (match and mismatch sets, differing in genres)[1] contain 5 annotations per item. The MNLI dev matched set contains 1,599 items for which exactly 3 annotators (out of the 5) agreed on the label. This subset was reannotated by Nie et al. (2020) with 100 annotations per item, called the ChaosNLI dataset. We randomly sampled 450 items from ChaosNLI. Figure 1 shows the annotations, with items organized by which label was the most frequent. While some items can be seen as having a unique ground truth label (depending on how many annotators agreeing on the same label are needed for that—here we take 80%, following Jiang and de Marneffe [2019]), other items clearly lead to differing annotations.

We also sampled 60 items from the MNLI dev matched set in which at most 2 out of the 5 annotators agreed on the label, and there are thus no majority labels. These items, coded with label ''–'' in Williams et al.'s (2018) release, are customarily discarded in evaluating NLI models.

3.2 Disagreement Categories

Our taxonomy of potential disagreement sources consists of 10 categories, shown in Table 1. The categories are organized into three high-level classes, corresponding to the three components of the annotation process in the Triangle of Reference: (1) uncertainty in the sentence meaning, (2) underspecification of the guidelines, and (3) annotator behavior.

3.2.1 Uncertainty in Sentence Meaning

Some textual phenomena leading to disagreement can be local to **Lexical** items, where the truth of the hypothesis depends on the meaning of a specific lexical item. That lexical item can have multiple meanings, or its meaning requires certain parameters that remain underspecified in the sentence at hand, as we saw with *many* in (1).[2] Disagreement can come from a pair of lexical items, where the lexical relationship between the items (e.g., hypernymy, synonymy) is loose, as in [1] in Table 1: Do people infer *advances in electronics* from *technological advances*?

Other cases involve the holistic meaning of the sentences and interpreting them in different contexts. In some cases, the hypothesis is an **Implicature** of the premise, as in [2]. By definition, an implicature can be canceled (Grice, 1975), which leads to a potential for differences in the annotations. Here, *some of the most authentic papyrus (are sold in The Pharaonic Village)* gives rise to the scalar implicature *but not all of the most authentic papyrus*, making the hypothesis false since it asserts that authentic papyrus are only sold in The Pharaonic Village. However, if the implicature is cancelled, *some* can also be interpreted as *all* (e.g., *Some students came. In fact, all came.*)

The hypothesis can target what is being **presupposed** by the premise. Wh-questions, for instance, presuppose that the entity the question

[1]The match dev set is of the same genre as the training set (e.g., fiction, government websites), while the mismatch dev set comes from other genres than the training set (e.g., face-to-face conversations, letters).

[2]It is challenging to distinguish between multiple senses or implicit parameters. For instance, in the pair *P: Then he sobered. - H: He was drunk.*, whether H can be inferred from P depends on the word *sober*: One could be sober from alcohol or from other drugs. Are these two meanings of the word or is the substance an implicit parameter?

bears on exists. The question *What changed?* in [3] presupposes that something changed, hence the answer *Nothing changed* can be viewed as contradictory. However, the premise can also be viewed as not giving enough information to judge the truth of the hypothesis, which would lead to a Neutral label.

Probabilistic Enrichment items involve making probabilistic inferences from the premise: The inferred content is likely, but not definitely, true in some contexts. In [4], there is some likelihood that nodding to the speaker's assertion means that one agrees with it. If annotators make that inference, they see the hypothesis as Entailment. But, since the premise is not explicitly stating the hypothesis, a Neutral label is also warranted.

Some premises/hypotheses contain typos or are fragments, making it hard to grasp their exact meaning (as in [5]). We call these cases **Imperfection**, following Williams et al. (2022).

3.2.2 Underspecification in the Guidelines

Some disagreements stem from the loose definition of the NLI task. Assuming **coreference** between the premise and the hypothesis has been noted as an important aspect of the NLI task (Mirkin et al., 2010) and necessary for obtaining high agreement in annotation (de Marneffe et al., 2008; Bowman et al., 2015; Kalouli et al., 2019). In [6], 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.

While the NLI task assumes coreference between entities and events mentioned in the premise and hypothesis, which entity/event to take into consideration is not always clear. For example, in (2), the premise can be taken to talk about ''desegregation being undone in Charlotte by magnet schools'', in which case the hypothesis is inferred.

(2) **P:** Unfortunately, the magnet schools began the undoing of desegregation in Charlotte.

H: Desegregation was becoming disbanded in Charlotte thanks to the magnet schools. [E,N,C]: [81, 6, 13]

The premise can also be taken to focus on the fact that ''the desegregation being undone in Charlotte by magnet schools is unfortunate''. In other words, two different ''Questions Under Discussion'' (Roberts, 2012) can be posited for the premise. Under that second interpretation, the hypothesis (in which the undoing of desegregation is positive, given the word *thanks*) contradicts the premise, where the desegregation undoing is unfortunate.

The truth of the hypothesis can also depend on the time at which the hypothesis is evaluated (**Temporal Reference**), but the NLI annotation guidelines do not specify how to handle such cases. There are two contextually salient temporal referents in [7], before or after the product release is issued. If the hypothesis refers to the time after the release is issued, it is true. From the perspective of before the release is issued, it is unclear whether the co-requesters can restrict timing or not.

Unlike assertions, questions do not have truth values (Groenendijk and Stokhof, 1984; Roberts, 2012). It is therefore theoretically ill-defined to ask whether an **interrogative hypothesis** is true or not given the premise (which is the question asked in Nie et al.'s (2020) annotation interface to build ChaosNLI). However, most of the interrogative hypotheses have interrogative premises (81.8% in MNLI dev sets; all in our subset). Groenendijk and Stokhof (1984) define the notion of entailment between questions: An interrogative *q1* entails another *q2* iff every proposition that answers *q1* answers *q2* as well. Some annotators seem to latch onto this definition, as in (3).

(3) **P:** yeah but uh do you have small kids

H: Do you have any children? [E,N,C]: [65,33,2]

Still, there is no definition distinguishing neutral from contradictory pairs of questions.[3] Annotators, perhaps to assign some meaning to the Neutral/Contradiction distinction, give judgments that seem to involve applying surface-level features for declarative sentences, choosing Neutral/Contradiction if the sentences involve substitution of unrelated words, as in (4).

(4) **P:** Where is art?

H: What is the place of virtue? [E,N,C]: [1,59,40]

[3]The issue does not necessarily arise from interrogative premises, since the hypothesis may target the presupposition of the question, as in [3].

3.2.3 Annotator Behavior

By definition, disagreement arises when a proportion of annotators behave one way and another proportion another way. We identified two patterns of ''systematic behavior'' (while it is hard to say for certain what annotators have in mind, the patterns seem robust). When the hypothesis adds content that provides minimal information compared to the premise, but is otherwise entailed, annotators are more likely to judge it as Entailment, thus ignoring/**accommodating minimally added content**. For instance, the hypothesis in [9] adds the information source (*said in the report*) which is not mentioned in the premise. From a strict semantic evaluation, the hypothesis is thus not inferred from the premise. Nonetheless most people are happy to infer it. Such added contents are often not at issue, that is, not the main point of utterance (Potts, 2005; Simons et al., 2010), appearing as modifiers (McNally, 2016), or parentheticals, making it easier for people to ignore if not paying enough attention or not being tuned to such differences.[4]

These biases are potentially problematic for applications in NLG that use the NLI labels for evaluating paraphrases (modeled as bi-directional entailment, [Sekine et al., 2007]), dialog coherence (Dziri et al., 2019), semantic accuracy (Dušek and Kasner, 2020), or use NLI as a pretraining task for learned metrics (Sellam et al., 2020). For instance, it would not be semantically accurate for a generated summary to hallucinate and include extraneous, even if not at-issue content, such as *said the report* in [9], if not already given in the source text.

When the hypothesis has **high lexical overlap** with the premise (e.g., involve the same noun phrases), annotators tend to judge it as Entailment even if it is not strictly inferred from the premise. In [10], the hypothesis claims that the white townsfolk thinks it sounds convincing, whereas the premise only states that the white townsfolk makes it sound convincing (and does not mention whose opinion it is). McCoy et al. (2019) pointed out that items in MNLI with high lexical overlap between the premise and the hypothesis often have the Entailment label, and that NLI models learn such shallow heuristics, ending up to incorrectly predict Entailment for items with high overlap. McCoy et al.'s (2019) finding might partially be attributed to such annotator behavior.

[4]Items belonging to other categories may also exhibit such pattern, such as [6] for which 24/100 annotators chose Entailment. Note that annotators for ChaosNLI were carefully vetted and passed multiple screening and training rounds.

3.3 Taxonomy Development and Annotation

The taxonomy was developed by a single annotator, starting by examining lowest and highest agreement examples in ChaosNLI to identify linguistic phenomena that are potential sources of disagreement in the NLI annotations. Some categories were merged because the distinction between them seem murky (for instance, the distinction of multiple senses vs. implicit argument in the Lexical category). Event coreference often requires entity coreference and the distinction between both is not clear-cut. For the two sentences *vendors crammed the streets with shrine offerings* and *vendors are lining the streets with torches and fires* to refer to the same event, we need to assume that they talk about the same set of vendors. We thus only have one Coreference category.

There were two rounds of annotations. In Round 1, one annotator annotated 400 items from ChaosNLI and iteratively refined the taxonomy, while writing annotation guidelines. Another annotator was then trained. In Round 2, both annotators annotated 50 additional items from ChaosNLI and 60 items from MNLI where only 2 out of the 5 original annotations agreed. These 110 items serve to check that the taxonomy does not ''overfit'' the 400-item sample used while developing it.

Multi-category Annotations More than one reason for disagreement may apply. We therefore adopt a multi-category annotation scheme: Each item can have multiple categories. For example, in (5), both Implicature and Temporal Reference contribute to disagreement. The premise does not suggest that the park changed name, while the hypothesis does so with the implicature triggered by *used to*. Therefore, if we evaluate the truth of the hypothesis now, there can be disagreement between Neutral and Contradiction. If we evaluate the truth of the hypothesis in or before 1935, the hypothesis is entailed because the park was named after Corbett at some point. Also, given that the implicature is triggered by a specific lexical

	Round 1		Round 2	
	#	%	#	%
Probabilistic	113	29.05	31	28.18
Lexical	65	16.71	32	29.09
Coreference	48	12.34	14	12.73
Accommodating	46	11.83	5	4.55
Imperfection	18	4.63	4	3.64
Lexical, Probabilistic	17	4.37	1	0.91
Interrogative.	15	3.86	5	4.55
Implicature, Lexical	11	2.83	0	0.00
Implicature	8	2.06	0	0.00
Coreference, Probabilistic	8	2.06	4	3.64
High Overlap	7	1.80	1	0.91
Presupposition	6	1.54	3	2.73
Temporal	5	1.29	2	1.82
Coreference, Lexical	5	1.29	0	0.00
Lexical, Presupposition	3	0.77	0	0.00
Implicature, Probabilistic	2	0.51	1	0.91
Coreference, Imperfection	1	0.26	2	1.82
Coreference, Temporal	1	0.26	1	0.91
Lexical, Temporal	1	0.26	1	0.91
Probabilistic, Temporal	1	0.26	1	0.91
Sub-Total	381	97.98	108	98.21

8 combinations occurred once only in Round 1
Accommodating, Probabilistic | Presupposition, Probabilistic | Lexical, Presupposition, Probabilistic | Accommodating, Lexical, Probabilistic | Imperfection, Lexical | Accommodating, Lexical | Coreference, Implicature | Implicature, Temporal

2 combinations occurred once only in Round 2
Presupposition, Temporal | High Overlap, Lexical, Probabilistic

Table 2: Frequency and percentage of each combination of categories in the taxonomy, in the two annotation rounds.

item (in contrast to non-conventional conversational implicatures), the category Lexical applies, too.

(5) **P:** The park was established in 1935 and was given Corbett's name after India became independent.

H: The park used to be named after Corbett.
[E,N,C]: [36, 34, 30]

Inter-annotator Agreement Since the annotation requires the understanding of various linguistic phenomena, only expert annotation is possible. The two annotators have graduate linguistic training. The Krippendorff's α with MASI distance (Passonneau, 2006) is 0.69. For the items annotated by both annotators, we then aggregated the two sets of annotations by taking their intersection. This resulted in 24 instances of categories deleted for annotator 1 (in 23 items) and 16 (in 16 items) for annotator 2. There were only 4 items (out of 110) with an empty intersection, which we reconciled.

Distribution of Categories Table 2 shows the frequency of each combination of categories for the two rounds of annotations. Probabilistic Enrichment and Lexical are the two most frequent categories, because they are broad categories by definition and not tied to a close set of lexical items. It also shows that the Round 2 annotations have roughly the same distribution, although-some combinations are rarer/did not appear in Round 1. No items have been encountered in Round 2 that needed creation of a novel disagreement category.

There are 11 items in the Round 1 sample for which none of the annotators could identify a source of disagreement. These items exhibit characteristics of clear, easily identifiable cases: paraphrase or containment relations, for Entailment (6); antonym or negation, for Contradiction (7); statements containing information that is not given by nor can be inferred from the premise, for Neutral (8). They also involve high agreement in the ChaosNLI annotations (average number of majority votes between 65 and 95, with a mean of 84.1).

(6) **P:** I'm confused.

H: Not all of it is very clear to me.
[E,N,C]: [92,5,3]

(7) **P:** uh-huh you can't do that in a skirt poor thing

H: You can do anything in a skirt.
[E,N,C]: [3, 23, 74]

(8) **P:** She had the pathetic aggression of a wife or mother–to Bunt there was no difference.

H: Bunt was raised motherless in an orphanage.
[E,N,C]: [0,88,12]

3.4 Findings and Discussion

Through the construction of the taxonomy, we found that disagreement arises from many reasons. The NLI annotations do not always show the full picture in terms of the range and nature of the meaning the sentences carry, because (1) even if an item has multiple possible interpretations, the annotators may converge on one of them, (2) there are at least two interpretations of the label distribution, arising out of a single probabilistic inference, or multiple categorical inferences.

Annotators Converge on One Interpretation NLI annotations for items exhibiting some of the factors that contribute to disagreement may actually show high agreement. Indeed, even when an

	Converge %	Total #	Mean (std) majority vote
Lexical	17.74	124	66.02 (14.15)
Implicature	12.50	24	63.96 (14.21)
Presupposition	0.00	12	57.92 (13.82)
Probabilistic Enrichment	13.33	165	64.56 (12.06)
Imperfection	22.73	22	67.18 (14.71)
Coreference	14.67	75	66.17 (14.39)
Temporal Reference	25.00	12	62.0 (19.33)
Interrogative Hypothesis	20.00	15	63.13 (14.32)
Accommodating	25.49	51	67.76 (16.48)
High Overlap	0.00	8	65.12 (4.75)

Table 3: For each disagreement category, the percentage of items exhibiting convergence (at least 80/100 annotators agreed on the same NLI label), the total number of items in the category, and the mean/standard deviation of the majority vote count.

item lends itself to uncertainty or multiple interpretations, a high proportion of annotators may converge to the same interpretation. For instance, in [1] (Table 1), 82 annotators (out of 100) take *technological advancement* to entail *advancement in electronics*, even though there are other kinds of technological advancement that are not electronics. In [5], 90 annotators latch onto the fact that the hypothesis seem totally unrelated to the premise, agreeing on the Neutral label.

Table 3 shows the percentage of items in each taxonomy category for which at least 80 (out of 100) annotators agreed on the same NLI label (which we will refer to as ''convergence''). Interestingly, ''Accommodating minimally added content'' has the largest amount of convergence (25.5%) and the highest mean majority vote (67.8). The majority voted labels are Entailment (accommodating the content) or Neutral (considering that the content is not given by the premise). Whether accommodation takes place depends on the extent to which the added content is not-at-issue and on the content itself. In [9] (Table 1), 97 annotators accommodated the extra content (*said the report*) in the hypothesis. In (9), however, the hypothesis also introduces new content *all year round*, but only 7 annotators accommodate it. In (10), 32 annotators accommodate the added content *American*, thus more than in (9) but less than in [9].

(9) **P:** The equipment you need for windsurfing can be hired from the beaches at Tel Aviv (marina), Netanya, Haifa, Tiberias, and Eilat.

H: Windsurfing equipment is available for hire in Tel Aviv all year round. [E,N,C]: [7, 93, 0]

(10) **P:** And here, current history adds a major point.

H: Current American history adds a major point. [E,N,C]: [32, 67, 1]

The difference could be due to the fact that *American* modifies the subject, which makes it less at-issue than *all year round* modifying the entire matrix clause. In [9], *said the report* appears in a parenthetical at the end of the sentence, which is even less at-issue than modifiers. Identifying such gradience in disagreement is a very difficult task: Simply identifying whether the hypothesis adds content is not enough, knowledge about the role of information structure seems necessary too.

Two Interpretations of NLI Label Distributions

It should now be clear that, by modeling majority vote, we are missing out on the full complexity of language understanding. Some argue that textual inference is probabilistic in nature (Glickman and Dagan, 2005). Therefore, probabilistic inferences give rise to disagreement in categorical labels (i.a., Zhang et al., 2021; Zhou et al., 2021). Here, we found that disagreement in the categorical labels arise in at least two ways: (1) a single probabilistic inference, or (2) multiple potentially categorical inferences, which is often the case when there are multiple possible specifications of the contextual factors (e.g., coreference, temporal reference, implicit arguments of some lexical items). They also differ in the kinds of uncertainty they exhibit. One is uncertainty in the state of the world. One is in how to interpret the sentences.

This distinction gives different interpretations of the aggregated label distribution. In [4] (Table 1), each annotator may have an underlying probabilistic judgment of how likely it is that *Sir James thinks it's absurd*, which is then reflected in the aggregated probability distribution. The probability associated with the Entailment label can be taken as the probabilistic belief (Kyburg, 1968) of an individual annotator for the truth of the hypothesis.

On the other hand, [7] involves categorical and probabilistic inferences. Whether the hypothesis is entailed depends on whether it is evaluated

before or after the product release is issued. If after, readers have a **categorical** judgment that the hypothesis is entailed. If before, readers have a **probabilistic** judgment, leading to the uncertainty between Neutral and Contradiction. Therefore, unlike [4], the probability associated with the Entailment label does not represent the judgment of an individual.

We could design experiments to collect empirical evidence for this distinction, such as collecting multilabel or sliding bar annotations, or free-text explanations to gain direct evidence of whether annotators have categorical/probabilistic judgments. Pursuing this line of research is left for future work.

Artificial Task Setup One of the reasons for the occurrence of disagreement may be the somewhat artificial setup of the NLI task. The premise and hypothesis are interpreted in isolation with no surrounding discourse. However, discourse context is needed for resolving much of the uncertainty in meaning pointed out here (e.g., coreference, temporal reference, and implicit arguments of lexical items). Investigating whether incorporating context into NLI annotations improves agreement is left for future work.

4 Modeling Experiments

Now that we understand better how disagreement arises, we explore how to build models that provide disagreement information. As discussed in Section 2, a distribution gives the most fine-grained information but can be misleading to interpret, while categorical information is often needed in downstream applications. Therefore, we experiment with models that provide two kinds of categorical information for disagreement: an additional ''Complicated'' class for labeling low agreement items (Section 4.3), and a multilabel classification approach, where each item is associated with one or more of the three standard NLI labels (Section 4.4). As baseline, we take the MixUp approach in Zhang et al. (2021), which predicts a distribution over the three labels, and uses a threshold to obtain multilabels/4-way labels.

These models can be useful in an annotation pipeline. One needs to collect multiple judgments for each item to cover the range of possible interpretations, but doing so may be prohibitively expensive at a large scale. The annotation budget could thus be prioritized by collecting annotations for items with potential for disagreement, as predicted by the model. Therefore, our goal is not necessarily to maximize accuracy. A model that can recall the possible interpretations is preferred to a model that misses them.

Dataset	E	N	C	EN	NC	EC	ENC
Chaos	195	57	37	291	205	32	76
				604 Complicated			
Chaos+Orig	1117	1117	1117	1117	775	163	76
				2131 Complicated			

Table 4: Number of items for each 4-way label and each combination of multilabel in each dataset. The number of ''Complicated'' items is the sum of the number of items with more than one label in the multilabel setup.

4.1 Training Data

We saw that there is gradience in disagreement, but we start with clearly delineated data and only take items for which there is distinct (dis)agreement. We first focus on items from ChaosNLI since they have 100 annotations each, giving a clearer signal for (dis)agreement, discarding items where the majority vote is between 60 and 80 (given that it is unclear whether this counts a high or low agreement). However, this gives a highly class-imbalanced set in both schemes, as shown in the line for ''Chaos'' in Table 4, with fewer items in E/N/C than in the other classes.[5] Therefore, we augment the set with data from the original MNLI dev sets (where items have 5 annotations). We use the following criteria to relabel the data with the 4-way scheme (E, N, C, and Complicated) and the multilabel scheme:

- Items receive a single E, N, or C label (in the 4-way and multilabel schemes) if the majority vote label has more than 80 votes (out of 100 annotations) for the ChaosNLI items or if all 5 annotations agree for the MNLI items.
- ChaosNLI items are labeled as Complicated or as having multiple labels if the majority has less than 60 votes. For multilabel, a label is present if it has at least 20 votes (complement of 80 used for the single label items).

[5]Both the baseline and our models perform poorly when trained with the imbalanced set.

MNLI items are labeled as Complicated if two labels have at least 2 votes (e.g., [3,2,0] or [2,2,1]). For multilabel, a label is present if it has at least 2 votes.

We downsampled[6] MNLI items with one of the E/N/C labels to get a class-balanced set in the multilabel scheme. The resulting sizes are shown in Table 4, line ‘‘Chaos+Orig’’. We split the ‘‘Chaos+Orig’’ set into train/dev/test with sizes 2710/816/1956, respectively, stratified by labels.

4.2 Baseline

We use Zhang et al.’s (2021) MixUp model as baseline for both the 4-way and multilabel schemes. The MixUp model has the same architecture as fine-tuning RoBERTa for classification. During training, each training example is a linear interpolation of two randomly chosen training items, for both the input encodings and the soft-labels (the annotation distributions over E/N/C). We used Zhang et al.’s hyperparameters, with a learning rate of 1e-6 and an early stopping patience of 5 epochs. The model is trained with the data split described above by optimizing KL-divergence with soft-labels.

To evaluate, we convert each predicted distribution to a multilabel, taking any label assigned a probability of at least 0.2 to be present (same threshold we used for the data). The multilabel is then converted to a 4-way label: Complicated if more than one label is present; E, N, or C if it is the only label. Comparing the results from the MixUp model with the ones from our approach will tell whether optimizing for distributions (as done by the MixUp model) gives better predictions than training with categorical labels (as done by our approach), when evaluating with categorical labels.

4.3 Four-way Classification

We fine-tuned RoBERTa (Liu et al., 2019) on the train/dev set using the standard methods for classification. We used the initial learning rate of 1e-5, with learning rate decay by 0.8 times if dev F1 does not improve for two epochs. We trained for up to 30 epochs, with early stopping used if dev F1 does not improve for 10 epochs. We used `jiant v1` (Wang et al., 2019b) for our experiments.

[6]We also experimented without downsampling majority vote: It worsened the performance on identifying the Complicated class or items with multiple labels.

	Chaos+Orig		Chaos		Orig	
	MixUp	Our	MixUp	Our	MixUp	Our
Accuracy	63.50	67.26	47.44	58.97	66.55	68.84
Macro Precision	63.64	69.69	42.14	47.71	66.59	71.97
Macro Recall	68.75	67.78	57.60	54.29	70.31	69.46
Macro F1	65.34	68.59	43.65	49.02	67.82	70.41
Complicated F1	48.76	62.32	51.91	68.28	47.73	60.46
E F1	69.59	68.69	50.93	48.42	72.92	72.27
N F1	65.91	67.50	35.43	42.77	69.20	69.67
C F1	77.11	75.84	36.32	36.64	81.45	79.23

Table 5: Left: 4-way classification performance on the test set. Right: Performance on the two subsets of the test set, Chaos and Original MNLI. Darker color indicates higher performance.

		Prediction				
		E	N	C	Complicated	All
Gold	E	274	10	3	108	395
	N	3	257	5	130	395
	C	9	6	297	83	395
	Complicated	116	87	76	492	771
	All	402	360	381	813	1956

Table 6: Confusion matrix of the 4-way classification predictions from the initialization with the highest macro F1. Darker color indicates higher numbers.

Results Table 5 shows accuracy, macro F1, and F1 for each class. Each score is the average from three random initializations. The macro F1 of our model is 68.59% (vs. 65.34% for MixUp), which is on par with previous work (Zhang and de Marneffe, 2021), but with room for improvement. Our model generally outperforms the baseline, suggesting that training with categorical labels is beneficial for predicting categorical labels.

‘‘Complicated’’ Is Most Confused The model performs worse on the Complicated label as opposed to the other three NLI labels. This is consistent with Kenyon-Dean et al.’s (2018) observation: The Complicated class is hard to model, due to its heterogeneity, as we saw in Section 3.2. This is also shown in the confusion matrix in Table 6. Conversely, there are few errors among the three original NLI classes, which is partly due to the stringent threshold we used to identify items on which we take the majority vote.

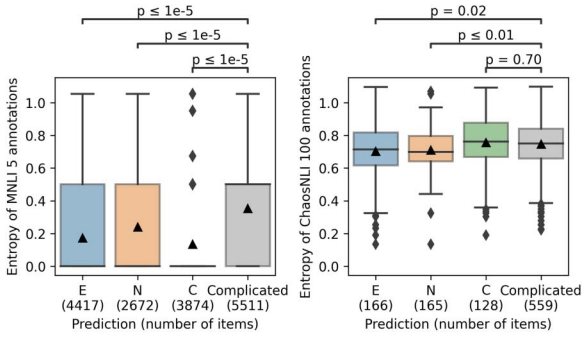

Figure 2: Boxplots of annotation entropy (Left: from original MNLI 5 annotations, Right: from ChaosNLI 100 annotations) by predicted label. Number of items shown in parentheses. Triangles indicate the means. P-values from Mann-Whitney two-sided test.

100 Annotations Are Better Since the Complicated label is the most confused, we investigate where the confusion comes from. We partition the test set by whether the label comes from the ChaosNLI 100 annotations or the original MNLI 5 annotations, and compare the Complicated F1 on each subset. Table 5 shows that for Complicated F1, the model performs much better on the Chaos subset with labels from 100 annotations than on the Original MNLI subset with labels from 5 annotations. This suggests that 100 annotations provide clearer training signals and are more informative as to whether the items exhibit disagreement.

Model Predicts ''Complicated'' on High Entropy Items We can see how the model performs on a larger scale, not just limiting to items where there is clear (dis)agreement. For this analysis, we get the model predictions on the full MNLI matched and mismatched dev sets, excluding the items used in our train and dev sets. We compare the model predictions with the annotation entropy, shown in Figure 2. Items predicted to be Complicated have significantly higher entropy than items predicted to be other labels (except for predicted Contradiction and Complicated items from ChaosNLI). This suggests that the model learned certain features associated with complicatedness.

4.4 Multilabel Classification

As mentioned in Section 2, the rationale for using a multilabel classification approach is to get insight in the way in which an item is complicated. Instead of choosing one of the three NLI labels (or four, including Complicated), the model is to predict multiple of the three NLI labels.

	Chaos+Orig		Chaos		Orig	
	MixUp	Our	MixUp	Our	MixUp	Our
Accuracy	58.59	59.44	38.14	44.98	62.47	62.19
Macro Precision	84.25	81.19	82.52	81.98	84.60	80.99
Macro Recall	76.81	83.35	66.65	77.35	79.41	84.92
Macro F1	80.29	82.17	73.62	79.48	81.79	82.76
1 Label Accuracy	77.86	69.23	57.66	45.05	79.95	71.73
2 Labels Accuracy	30.20	46.16	32.52	52.97	29.53	44.21
3 Labels Accuracy	5.26	10.53	5.26	10.53	0.00	0.00

Table 7: Left: Multilabel classification performance on the test set. Right: Performance on the two test set subsets. The Orig subset does not have any items with all three labels present.

Model Architecture To perform multilabel classification with 3 labels, we make minimal changes to the standard method for fine-tuning RoBERTa for 3-way classification. We predict each E/N/C label independently, by applying the sigmoid function on top of the 3 logits given by the MLP classifier on top of RoBERTa to obtain probabilities associated with each label. We take the label to be present if its probability is greater than 0.5. The model is trained with a cross entropy loss.

Training Procedure We used an initial learning rate of 5e-6, with learning rate decay by 0.8 times if dev F1 does not improve for one epoch. We trained for up to 30 epochs, with early stopping used if dev F1 does not improve for 10 epochs.

Results Table 7 gives the macro precision, recall, and F1, and the exact match accuracy partitioned by the number of gold labels (1/2/3 Labels Accuracy) for the test set and for its subsets. Our model has a higher F1 score than the baseline but a lower precision. The baseline model is more successful at items on which annotators agree (higher 1 Label Accuracy), while our model performs better on items with disagreement (2/3 Labels Accuracy).

Comparing the two test set subsets, we see the same pattern as in the 4-way results: On disagreement items, our model performs better (higher 2/3 Labels Accuracy) on the Chaos subset than on the Orig subset. This corroborates the finding from the 4-way classification that 100 annotations give a better indication of complicatedness.

	E	N	C	EN	NC	EC	ENC
E	3718	8	19	503	5	178	32
N	9	2076	4	467	275	0	8
C	7	6	3170	35	479	135	49
Complicated	395	838	300	1995	1289	162	312

Table 8: Contingency matrix of the 4-way classification vs. the multilabel predictions, on the full MNLI dev sets (excluding items used in our train/dev sets).

Multilabel Model Is More Expressive The accuracy decreases from the 4-way classification setup, which is expected since the number of possible labels increased from 4 to 7 (all possible combinations of the 3 labels). However, the macro recall increases compared to the 4-way classification (83.35 vs. 67.78), possibly as a result of more expressivity in the model output and not having one challenging and heterogeneous class. We also see this more concretely in the contingency table of the 4-way model vs. the multilabel model predictions (Table 8): When the multilabel model predicts more than one label, the 4-way model often predicts the Complicated class or one of the labels predicted by the multilabel model. In other words, the 4-way model may miss one or more labels while the multilabel model can identify all of them.

Takeaways Comparing with the MixUp baseline, which is trained with soft-labels, we see that training with categorical labels performs better in predicting categorical labels. Therefore, for downstream tasks where categorical information is needed, training with categorical labels is recommended. The multilabel model is more expressive, and as we will show in Section 5, it provides fine-grained information that gives a better understanding of what the model has learned. Our results suggest that the multilabel approach could potentially be used as intrinsic evaluation for how well the model captures the judgments of the population.

5 Error Analysis

We analyze the model behavior with respect to the categories of disagreement sources. For each category, Figure 3 gives the percentages of ChaosNLI items annotated with at least that category and having converging NLI interpretations

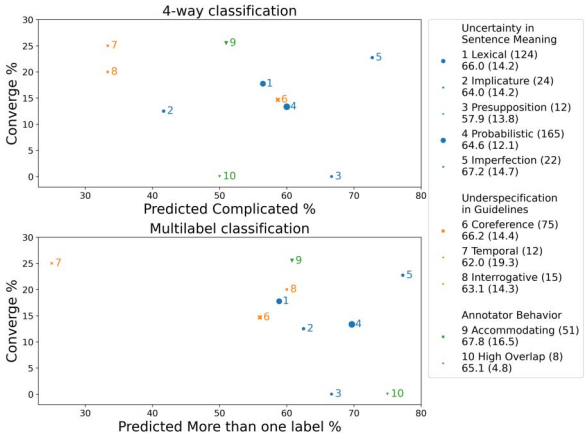

Figure 3: For each disagreement category, percentage of ChaosNLI items annotated with that category (number in parentheses) and having converging NLI annotations (>80 majority vote) vs. percentage predicted as Complicated in the 4-way setup or as having more than one label in the multilabel setup. Legend also gives mean majority vote in each category, with standard deviation in parentheses.

(>80 agree on the NLI label) vs. percentages of items predicted to exhibit disagreement (Complicated by the 4-way model or as having more than one label by the multilabel model). Overall, a category with more agreement (higher majority vote) in the annotations tends to have fewer items predicted as exhibiting disagreement. This is expected given that an item with convergence corresponds to not having disagreement as gold label, and the model performs well overall.

Comparing the two models, we see that all categories, except [7] Temporal Reference, are farther to the right in the Multilabel classification (bottom panel) whereas they are more spread out in the 4-way classification (top panel), meaning that the 4-way model predicts an agreement label (E/N/C) more often than the Multilabel model. This suggests that the 4-way model is more strongly tied to the convergence statistics and fails to detect potentials of disagreement. It also aligns with the previous finding that the Multilabel model has higher recall of the range of interpretations.

Items in [3] Presupposition, [4] Probabilistic Enrichment, and [5] Imperfection are often predicted in both setups to exhibit disagreement (they are to the right of both plots in Figure 3). [6] Coreference, [2], Implicature, and [10] High Overlap also appear to the right, depending on

	Premise	Hypothesis	[E,N,C]	Predictions
	Probabilistic Enrichment			
1	Oh, sorry, wrong church.	He or she entered the wrong church.	[82, 17, 1]	Complicated / EN
2	There should be someone here who knew more of what was going on in this world than he did now.	He knew things, but hoped someone else knew more.	[82, 18, 0]	Complicated / EN
3	What am I to do with them afterwards?"	It is the narrator's responsibility to take care of them.	[15, 73, 12]	N / NC
4	But they persevered, she said, firm and optimistic in their search, until they were finally allowed by a packed restaurant to eat their dinner off the floor.	Because all of the seats were stolen, they had to eat off the floor.	[23, 57, 20]	Complicated / NC
	Coreference			
5	The original wax models of the river gods are on display in the Civic Museum.	They have models made out of clay.	[5, 38, 57]	C / C
6	Indeed, said San'doro.	Indeed, they said.	[52, 22, 26]	E / EN
7	This was built 15 years earlier by Jahangir's wife, Nur Jahan, for her father, who served as Mughal Prime Minister.	Nur Jahan's husband Jahangir served as Mughal Prime Minister.	[17, 17, 66]	Complicated / E
8	Cruises are available from the Bhansi Ghat, which is near the CityPalace.	You can take cruises from Phoenix Arizona.	[0, 51, 49]	Complicated /NC
	Accommodating Minimally Added Content			
9	The key question may be not what Hillary knew but when she knew it.	According to current reports, the question is not if, but when did Hillary know about it.	[90, 9, 1]	E / EN
10	Four or five from the town rode past, routed by their diminished numbers and the fury of the Kal and Thorn.	Kal and Thorn were furious at the villagers.	[50, 41, 9]	N / EN

Table 9: Examples from the categorization with ChaosNLI annotations and 4-way/multi-label model predictions.

the setup. Among those categories, [3] Presupposition, [2] Implicature, [5] Imperfection, and [10] High Overlap are associated with surface patterns, potentially making it easier for the models to learn that they often exhibit disagreement. We thus take a closer look at the following categories, across all items annotated: Probabilistic Enrichment, Coreference, and Accommodating Minimally Added Content (discussed in Section 3.4).

Probabilistic Enrichment The multilabel model predicts 68% of the items annotated with Probabilistic Enrichment to have more than one NLI label. In particular, 36% are predicted as EN and 27% as NC, corresponding the common patterns in Probabilistic Enrichment where the enriched (not explicitly stated) inference leads to Entailment/Contradiction, and the Neutral label is warranted without enrichment. We found that the multilabel model often predicts labels when they are only slightly below the threshold of 20 that we used to count a label as present (items 1 and 2 in Table 9). Even though in those cases the model is ''incorrect'' when calculating the metrics, it shows that the model can retrieve subtle inferences. In item 1, 17 annotators chose Neutral, while 82 chose Entailment: the premise does not mention *entering* a church, but most annotators take that situation to be likely. The multilabel model is, however, predicting both Entailment and Neutral, accounting for the possible interpretations.

Coreference For items annotated with Coreference, both models predict Entailment/Contradiction when the premise and hypothesis share the same argument structure or involve simple word substitutions (e.g., *wax/clay* in item 5 and *San'doro/they* in item 6, Table 9), which are features of unanimous Entailment/Contradiction. This suggests that such predictions are influenced by the unanimous items. The 4-way model tends to predict Complicated when items annotated with Coreference do not share any structure (as in items 7 and 8).

Accommodating Minimally Added Content The multilabel model predicts 44% of the items involving minimally added content to have both Entailment and Neutral labels, and 76% of the items to have at least the Neutral label. This is consistent with the majority of these items showing disagreement over Entailment and Neutral, and the sentences themselves exhibiting features of Neutral (added content) and surface features of Entailment (high lexical overlap), as in items 9 and 10. In item 9, the multilabel model recovers a Neutral inference (the premise does not mention current reports), even when only 9 annotators chose the Neutral label. This further illustrates that the multilabel model is better at recalling possible interpretations.

6 Conclusion

We examined why disagreement in NLI annotations occurs and found that it arises out of all three components of the annotation process. We experimented with modeling NLI disagreement as 4-way and multilabel classifications, and showed

that the multilabel model gives a better recall of the range of interpretations. We hope our findings will shed light on how to improve the NLI annotation process, for example, ways to specify the guidelines to reduce disagreement or introduce contexts that resolve underspecification, ways to gather enough annotations to cover the possible interpretations, as well as ways to model NLI without the single ground truth assumption.

Acknowledgments

We thank TACL editorial assistant Cindy Robinson and action editor Anette Frank for the time they committed to the review process. We thank Anette Frank for her clear and detailed editor letter, as well as the anonymous reviewers for their insightful feedback. We also thank Micha Elsner and Michael White, and members of the OSU Pragmatics and Clippers discussion groups for their suggestions and comments, and especially Angélica Aviles Bosques for her help with the annotation. This material is based upon work supported by the National Science Foundation under grant no. IIS-1845122. Marie-Catherine de Marneffe is a Research Associate of the Fonds de la Recherche Scientifique – FNRS.

References

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15. https://doi.org/10.1609/aimag.v36i1.2564

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1075

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110. https://doi.org/10.1162/tacl_a_00449

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 146–148, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1381

Katrin Erk and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 440–449, Singapore. Association for Computational Linguistics.

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2021. Did they answer? Subjective acts and intents in conversational discourse. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1626–1644, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.129

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.204

Oren Glickman and Ido Dagan. 2005. A probabilistic setting and lexical coocurrence model for textual entailment. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 43–48, Ann Arbor, Michigan. Association for Computational Linguistics. https://doi.org/10.3115/1631862.1631870

Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3411764.3445423

Herbert P. Grice. 1975. Logic and conversation, *Speech Acts*, pages 41–58. Brill. https://doi.org/10.1163/9789004368811_003

Jeroen Antonius Gerardus Groenendijk and Martin Johan Bastiaan Stokhof. 1984. *Studies on the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, University of Amsterdam.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. Evaluating BERT for natural language inference: A case study on the CommitmentBank. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6086–6091, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1630

Aikaterini-Lida Kalouli, Annebeth Buis, Livy Real, Martha Palmer, and Valeria de Paiva. 2019. Explaining simple natural language inference. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 132–143, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4016

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1171

Henry E. Kyburg. 1968. Bets and beliefs. *American Philosophical Quarterly*, 5(1):54–63.

John P. Lalor, Hao Wu, and Hong Yu. 2017. Soft label memorization-generalization for natural language inference.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach.

Peter LoBue and Alexander Yates. 2011. Types of common-sense knowledge needed for recognizing textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 329–334, Portland, Oregon, USA. Association for Computational Linguistics.

Jackson Luken, Nanjiang Jiang, and Marie-Catherine de Marneffe. 2018. QED: A fact verification system for the FEVER shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 156–160, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5526

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*,

38:301–333. https://doi.org/10.1162/COLI_a_00097

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pages 1039–1047, Columbus, Ohio. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*, 23(2):107–124.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1334

Louise McNally. 2016. Modification, Maria Aloni and Paul Dekker, editors, *The Cambridge Handbook of Formal Semantics*, Cambridge Handbooks in Language and Linguistics, chapter 15, pages 442–464. Cambridge University Press.

Shachar Mirkin, Ido Dagan, and Sebastian Padó. 2010. Assessing the role of discourse references in entailment inference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1209–1219, Uppsala, Sweden. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.734

Byung-Doh Oh, Pranav Maneriker, and Nanjiang Jiang. 2019. THOMAS: The hegemonic OSU morphological analyzer using seq2seq. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 80–86, Florence, Italy. Association for Computational Linguistics.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy. European Language Resources Association (ELRA).

Rebecca J. Passonneau, Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. 2012. Multiplicity and word sense: Evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46(2):219–252. https://doi.org/10.1007/s10579-012-9188-x

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694. https://doi.org/10.1162/tacl_a_00293

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014a. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden. Association for Computational Linguistics. https://doi.org/10.3115/v1/E14-1078

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014b. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-2083

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 76–83, Ann Arbor, Michigan. Association for Computational Linguistics. https://doi.org/10.3115/1608829.1608840

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of

multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1778–1789, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1176

Christopher Potts. 2005. *The Logic of Conventional Implicatures*. Oxford Studies in Theoretical Linguistics. Oxford University Press, Oxford.

Marta Recasens, Eduard Hovy, and M. Antònia Martí. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152. https://doi.org/10.1016/j.lingua.2011.02.004

Craige Roberts. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5. https://doi.org/10.3765/sp.5.6

Mark Sammons, V. G. Vinod Vydiswaran, and Dan Roth. 2010. ''Ask not what textual entailment can do for you...''. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1199–1208, Uppsala, Sweden. Association for Computational Linguistics.

Satoshi Sekine, Kentaro Inui, Ido Dagan, Bill Dolan, Danilo Giampiccolo, and Bernardo Magnini, editors. 2007. *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. Prague. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.704

Mandy Simons, Judith Tonhauser, David Beaver, and Craige Roberts. 2010. What projects and why. In *Semantics and linguistic theory*, volume 20, pages 309–327. https://doi.org/10.3765/salt.v20i0.2584

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.704

Alexandra Uma, Dina Almanea, and Massimo Poesio. 2022. Scaling and disagreements: Bias, noise, and ambiguity. *Frontiers in Artificial Intelligence*, 5. https://doi.org/10.3389/frai.2022.818451, PubMed: 35434607

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470. https://doi.org/10.1613/jair.1.12752

Yannick Versley. 2008. Vagueness and referential ambiguity in a large-scale annotated corpus. *Research on Language and Computation*, 6(3):333–353. https://doi.org/10.1007/s11168-008-9059-1

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*. https://doi.org/10.18653/v1/W18-5446

Alex Wang, Ian F. Tenney, Yada Pruksachatkun, Phil Yeres, Jason Phang, Haokun Liu, Phu Mon Htut, Katherin Yu, Jan Hula, Patrick Xia, Raghu Pappagari, Shuning Jin, R. Thomas McCoy, Roma Patel, Yinghui Huang, Edouard Grave, Najoung Kim, Thibault Févry, Berlin Chen, Nikita Nangia, Anhad Mohananey, Katharina Kann, Shikha Bordia, Nicolas Patry, David Benton, Ellie Pavlick, and Samuel R. Bowman. 2019b. jiant 1.3: A software toolkit for research on general-purpose text understanding models. http://jiant.info/.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge

corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1101

Adina Williams, Tristan Thrush, and Douwe Kiela. 2022. ANLIzing the adversarial natural language inference dataset. In *Proceedings of the Society for Computation in Linguistics*, volume 5. University of Massachusetts Amherst.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Learning with different amounts of annotation: From zero to many labels. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7620–7632, Online and Punta Cana, Dominican Republic, Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.601

Xinliang Frederick Zhang and Marie-Catherine de Marneffe. 2021. Identifying inherent disagreement in natural language inference. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4908–4915, Online. Association for Computational Linguistics.

Xiang Zhou, Yixin Nie, and Mohit Bansal. 2021. Distributed NLI: learning to predict human opinion distributions for language reasoning. *CoRR*, abs/2104.08676. https://doi.org/10.18653/v1/2022.findings-acl.79