Identifying inherent disagreement in natural language inference

Xinliang Frederick Zhang
Department of Computer Science and Engineering
The Ohio State University
zhang.9975@osu.edu

Marie-Catherine de Marneffe
Department of Linguistics
The Ohio State University
demarneffe.1@osu.edu

Abstract

Natural language inference (NLI) is the task of determining whether a piece of text is entailed, contradicted by or unrelated to another piece of text. In this paper, we investigate how to tease systematic inferences (i.e., items for which people agree on the NLI label) apart from disagreement items (i.e., items which lead to different annotations), which most prior work has overlooked. To distinguish systematic inferences from disagreement items, we propose Artificial Annotators (AAs) to simulate the uncertainty in the annotation process by capturing the modes in annotations. Results on the CommitmentBank, a corpus of naturally occurring discourses in English, confirm that our approach performs statistically significantly better than all baselines. We further show that AAs learn linguistic patterns and context-dependent reasoning.

1 Introduction

Learning to effectively understand unstructured text is integral to Natural Language Understanding (NLU), covering a wide range of tasks such as question answering, semantic textual similarity and sentiment analysis. Natural language inference (NLI), an increasingly important benchmark task for NLU research, is the task of determining whether a piece of text is entailed, contradicted by or unrelated to another piece of text (i.a., Dagan et al., 2005; MacCartney and Manning, 2009).

Pavlick and Kwiatkowski (2019) observed inherent disagreements among annotators in several NLI datasets, which cannot be smoothed out by hiring more people. They pointed out that to achieve robust NLU, we need to be able to tease apart systematic inferences (i.e., items for which most people agree on the annotations) from items inherently leading to disagreement. The last example in Table 1, from the CommitmentBank (de Marneffe et al., 2019), is a typical disagreement item: some annotators consider it to be an entailment (3 or 2), while others view it as a contradiction (-3). A common practice to generate an inference label from annotations is to take the average (i.a., Pavlick and Callison-Burch, 2016). In this case, the average of the annotations is 0.25 and the gold label for this item would thus be "Neutral", but such label is not accurately capturing the annotation distribution. Alternatively, some work simply ignores items on which annotators disagree and only studies systematic inference items (Jiang and de Marneffe, 2019a,b; Raffel et al., 2019).

Here, we aim at teasing apart systematic inferences from inherent disagreements. In line with what Kenyon-Dean et al. (2018) suggested for sentiment analysis, we propose a finer-grained labeling

1	*Premise:* Some of them, like for instance the farm in Connecticut, are quite small. If I like a place I buy it. I guess you could say it's a hobby. *Hypothesis:* buying places is a hobby. `Entailment (Entailment)` [3, 3, 2, 2, 2, 2, 1, 1]
2	*Premise:* "I hope you are settling down and the cat is well." This was a lie. She did not hope the cat was well. *Hypothesis:* the cat was well. `Neutral (Neutral)` [0, 0, 0, 0, 0, 0, 0, 0, -3]
3	*Premise:* "All right, so it wasn't the bottle by the bed. What was it, then?" Cobalt shook his head which might have meant he didn't know or might have been admonishment for Oliver who was still holding the bottle of wine. *Hypothesis:* Cobalt didn't know. `Neutral (Disagreement)` [1, 0, 0, 0, 0, 0, 0, -2]
4	*Premise:* A: No, it doesn't. B: And, of course, your court system when you get into the appeals, I don't believe criminal is in a court by itself. *Hypothesis:* criminal is in a court by itself. `Contradiction (Contradiction)` [-1, -1, -2, -2, -2, -2, -2, -3]
5	*Premise:* A: The last one I saw was Dances With The Wolves. B: Yeah, we talked about that one too. And he said he didn't think it should have gotten all those awards. *Hypothesis:* Dances with the Wolves should have gotten all those awards. `Contradiction (Disagreement)` [0, 0, -1, -1, -2, -2, -2, -3]
6	*Premise:* Meg realized she'd been a complete fool. She could have said it differently. If she'd said Carolyn had borrowed a book from Clare and wanted to return it they 'd have given her the address. *Hypothesis:* Carolyn had borrowed a book from Clare. `Disagreement (Disagreement)` [3, 3, 3, 2, 0, -3, -3, -3]

Table 1: Examples from CommitmentBank, with finer-grained NLI labels. The labels in parentheses come from Jiang and de Marneffe (2019b). Scores in brackets are the raw human annotations.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4908–4915
June 6–11, 2021.

	Entailment	Neutral	Contradiction	Disagreement	Total
Train	177	57	196	410	840
Dev	23	9	22	66	120
Test	58	19	54	109	240
Total	258	85	272	585	1,200

Table 2: Number of items in each class in train/dev/test.

for NLI: teasing disagreement items, labeled "Disagreement", from systematic inferences, which can be "Contradiction", "Neutral" or "Entailment". To this end, we propose Artificial Annotators (AAs), an ensemble of BERT models (Devlin et al., 2019), which simulate the uncertainty in the annotation process by capturing modes in annotations. That is, we expect to utilize simulated modes of annotations to enhance finer-grained NLI label prediction.

Our results, on the CommitmentBank, show that AAs perform statistically significantly better than all baselines (including BERT baselines) by a large margin in terms of both F1 and accuracy. We also show that AAs manage to learn linguistic patterns and context-dependent reasoning.

2 Data: The CommitmentBank

The CommitmentBank (CB) is a corpus of 1,200 naturally occurring discourses originally collected from news articles, fiction and dialogues. Each discourse consists of up to 2 prior context sentences and 1 target sentence with a clause-embedding predicate under 4 embedding environments (negation, modal, question or antecedent of conditional). Annotators judged the extent to which the speaker/author of the sentences is committed to the truth of the content of the embedded clause (CC), responding on a Likert scale from +3 to -3, labeled at 3 points (+3/speaker is certain the CC is true, 0/speaker is not certain whether the CC is true or false, -3/speaker is certain the CC is false). Following Jiang and de Marneffe (2019b), we recast CB by taking the context and target as the premise and the embedded clause in the target as the hypothesis.

Common NLI benchmark datasets are SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018), but these datasets have only one annotation per item in the training set. CB has at least 8 annotations per item, which permits to identify items on which annotators disagree. Jiang and de Marneffe (2019b) discarded items if less than 80% of the annotations are within one of the following three ranges: [1,3] Entailment, 0 Neutral, [-3,-1] Contradiction. The gold label for example 3 in Table 1 would thus be "Disagreement". However, this seems a bit too stringent, given that 70% of the annotators all agree on the 0 label and there is only one annotation towards the extreme. Likewise, for example 5, most annotators chose a negative score and the item might therefore be better labeled as "Contradiction" rather than "Disagreement". To decide on the **finer-grained NLI labels**, we therefore also took variance and mean into account, as follows:[1]

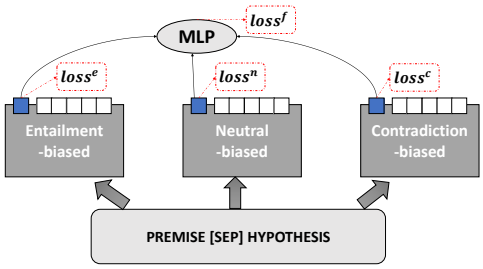

Figure 1: Artificial Annotators setup.

- **Entailment:** 80% of annotations fall in the range [1,3] OR the annotation variance ≤ 1 and the annotation mean > 1.
- **Neutral:** 80% of annotations is 0 OR the annotation variance ≤ 1 and the absolute mean of annotations is bound within 0.5.
- **Contradiction:** 80% of annotations fall in the range [-3, -1] OR the annotation variance ≤ 1 and the annotation mean < -1.
- **Disagreement:** Items which do not fall in any of the three categories above.

We randomly split CB into train/dev/test sets in a 7:1:2 ratio.[2] Table 2 gives splits' basic statistics.

3 Model: Artificial Annotators

We aim at finding an effective way to tease items leading to systematic inferences apart from items leading to disagreement. As pointed out by Calma and Sick (2017), annotated labels are subject to uncertainty. Annotations are indeed influenced by several factors: workers' past experience and concentration level, cognition complexities of items, etc. They proposed to simulate the annotation process in an active learning paradigm to make use of the annotations that contribute to uncertainty. Likewise, for NLI, Gantt et al. (2020) observed that directly training on raw annotations using annotator

[1] Compared with the labeling scheme in Jiang and de Marneffe (2019b), our labeling scheme results in 59 fewer Disagreement items, 48 of which are labeled as Neutral.

[2] We don't follow the SuperGLUE splits (Wang et al., 2019) as they do not include disagreement items. The data splits and codes are available at https://github.com/FrederickXZhang/FgNLI.

	Dev		Test					
	Acc.	F1	Acc.	F1	Entail	Neutral	Contradict	Disagree
Always 0	55.00	39.03	45.42	28.37	0.00	0.00	0.00	62.46
CBOW	55.25	40.54	45.09	28.37	0.00	0.00	0.69	62.17
Heuristic	65.00	62.08	54.17	50.60	22.54	**52.94**	64.46	58.20
Vanilla BERT	63.71	63.54	62.50	61.93	59.26	49.64	69.09	61.93
Joint BERT	64.47	64.28	62.61	62.07	59.77	47.27	67.36	63.21
AAs (ours)	**65.15**	**64.41**	**65.60***	**64.97***	**61.07**	51.27	**70.89**	**66.49***

Table 3: Baselines and AAs overall performance on CB dev and test sets, and F1 scores of each class on the test set (average of 10 runs). * indicates a statistically significant difference (t-test, $p \leq 0.01$).

identifier improves performance. Essentially, Gantt et al. (2020) used a mixed-effect model to learn a mapping from an item and the associated annotator identifier to a NLI label. However, annotator identifiers are not always accessible, especially in many datasets that have been there for a while. Thus, we decide to simulate the annotation process instead of learning from real identifiers.

As shown by Pavlick and Kwiatkowski (2019), if annotations of an item follow unimodal distributions, then it is suitable to use aggregation (i.e., take an average) to obtain a inference label; but such an aggregation is not appropriate when annotations follow multi-modal distributions. Without loss of generality, we assume that items are associated with n-modal distributions, where $n \geq 1$. Usually, systematic inference items are tied to unimodal annotations while disagreement items are tied to multi-modal annotations. We, thus, introduce the notion of Artificial Annotators (AAs), where each individual "annotator" learns to model one mode.

3.1 Architecture

AAs is an ensemble of n BERT models (Devlin et al., 2019) with a primary goal of finer-grained NLI label prediction. n is determined to be 3 as there are up to 3 relationships between premise and hypothesis, excluding the disagreement class. Within AAs, each BERT is trained for an auxiliary systematic inference task which is to predict entailment/neutral/contradiction based on a respective subset of annotations. The subsets of annotations for the three BERT are mutually exclusive.

A high-level overview of AAs is shown in Figure 1. Intuitively, each BERT separately predicts a systematic inference label, each of which represents a mode[3] of the annotations. The representations of these three labels are further aggregated as augmented information to enhance final fine-grained NLI label prediction (see Eq. 1).

If we view the AAs as a committee of three members, our architecture is reminiscent of the Query by Committee (QBC) (Seung et al., 1992), an effective approach for active learning paradigm. The essence of QBC is to select unlabeled data for labeling on which disagreement among committee members (i.e., learners pre-trained on the same labeled data) occurs. The selected data will be labeled by an oracle (e.g., domain experts) and then used to further train the learners. Likewise, in our approach, each AA votes for an item independently. However, the purpose is to detect disagreements instead of using disagreements as a measure to select items for further annotations. Moreover, in our AAs, the three members are trained on three disjoint annotation partitions for each item (see Section 3.2).

3.2 Training

We first sort the annotations in descending order for each item and divide them into three partitions.[4] For each partition, we generate an auxiliary label derived from the annotation mean. If the mean is greater/smaller than +0.5/-0.5, then it's entailment/contradiction; otherwise, it's neutral. The first BERT model is always enforced to predict the auxiliary label of the first partition to simulate an entailment-biased annotator. Likewise, the second and third BERT models are trained to simulate neutral-biased and contradiction-biased annotators.

Each BERT produces a pooled representation for the `[CLS]` token. The three representations are passed through a multi-layer perceptron (MLP) to obtain the finer-grained NLI label:

$$P(y|\mathbf{x}) = \text{softmax}(\mathbf{W_s} \tanh(\mathbf{W_t}[\mathbf{e}; \mathbf{n}; \mathbf{c}])) \quad (1)$$

[3] It's possible that three modes collapse to (almost) a point.

[4] For example, if there are 8 annotations for a given item, the annotations are divided into partitions of size 3, 2 and 3.

with $[\mathbf{e}; \mathbf{n}; \mathbf{c}]$ being the concatenation of three learned representations out of **e**ntailment-biased, **n**eutral-biased and **c**ontradiction-biased BERT models. $\mathbf{W_s}$ and $\mathbf{W_t}$ are parameters to be learned.

The overall loss is defined as the weighted sums of four cross-entropy losses:

$$loss = r * loss^f + \frac{1-r}{3}(loss^e + loss^n + loss^c) \quad (2)$$

where $r \in [0, 1]$ controls the primary finer-grained NLI label prediction task loss ratio.

4 Experiment

We include five baselines to compare with:

- **"Always 0"**: Always predict Disagreement.
- **CBOW** (Continuous Bags of Words): Each item is represented as the average of its tokens' GLOVE vectors (Pennington et al., 2014).
- **Heuristic baseline**: Linguistics-driven rules (detailed in Appendix A), adapted from Jiang and de Marneffe (2019b); e.g., conditional environment discriminates for disagreement items.
- **Vanilla BERT**: (Devlin et al., 2019) Straightforwardly predict among 4 finer-grained NLI labels.
- **Joint BERT**: Two BERT models are jointly trained, each of which has a different speciality. The first one (2-way) identifies whether a sentence pair is a disagreement item. If not, this item is fed into the second BERT (3-way) which carries out systematic inference.

For all baselines involving BERT, we follow the standard practice of concatenating the premise and the hypothesis with `[SEP]`.

Table 3 gives the accuracy and F1 for each baseline and AAs, on the CB dev and test sets. We run each model 10 times, and report the average.

CBOW is essentially the same as the "Always 0" baseline as it keeps predicting Disagreement regardless of the input. The Heuristic baseline achieves competitive performance on the dev set, though it has a significantly worse result on the test set. Not surprisingly, both BERT-based baselines outperform the Heuristic on the test set: fine-tuning BERT often lead to better performance, including for NLI (Peters et al., 2019; McCoy et al., 2019). These observations are consistent with Jiang and de Marneffe (2019b) who observed a similar trend, though only on systematic inferences. Our proposed AAs perform consistently better than all baselines, and statistically significantly better on the test set (t-test, $p \leq 0.01$). Also, AAs achieve a smaller standard deviation on the test set within the 10 runs, indicating that it is more stable and potentially more robust to wild environments.

1	*Premise:* B: Yeah, it is. A: For instance, B: I'm a historian, and my father had kept them, I think, since nineteen twenty-seven uh, but he burned the ones from twenty-seven to fi-, A: My goodness. B: I could not believe he did that, *Hypothesis:* his father burned the ones from twenty-seven Heuristics: C V. BERT: D J. BERT: E AAs: E {E, E, E} Gold: E [3, 3, 3, 3, 3, 2, 2, -1]
2	*Premise:* 'She was about to tell him that was his own stupid fault and that she wasn't here to wait on him - particularly since he had proved to be so inhospitable. But she bit back the words. Perhaps if she made herself useful he might decide she could stay - for a while at least just until she got something else sorted out. *Hypothesis:* she could stay Heuristics: D V. BERT: D J. BERT: D AAs: N {N, N, N} Gold: N [3, 0, 0, 0, 0, 0, 0, 0, 0, 0]
3	*Premise:* A: but that is one of my solutions. Uh... B: I know here in Dallas that they have just instituted in the last couple of years, uh, a real long period of time that you can absentee vote before the elections. And I do not think they have seen a really high improvement. *Hypothesis:* they have seen a really high improvement. Heuristics: C V. BERT: C J. BERT: C AAs: C {C, C, C} Gold: C [-1, -2, -2, -2, -2, -2, -2, -2, -3, -3]
4	*Premise:*B: So did you commute everyday then or, A: No. B: Oh, okay. A: No, no, it was a six hour drive. B: Oh, okay, when you said it was quite a way away, I did not know that meant you had to drive like an hour *Hypothesis:* speaker A had to drive like an hour Heuristics: C V. BERT: D J. BERT: E AAs: D {E, C, C} Gold: D [3, 2, 2, 1, 0, 0, -1, -1, -1, -3]
5	*Premise:* The assassin's tone and bearing were completely confident. If he noticed that Zukov was now edging further to the side widening the arc of fire he did not appear to be troubled. *Hypothesis:* Zukov was edging further to the side Heuristics: D V. BERT: D J. BERT: D AAs: D {E, E, N} Gold: E [3, 3, 3, 3, 2, 2, 1, 1]
6	*Premise:* B: Yeah, and EDS is very particular about this, hair cuts, A: Wow. B: I mean it was like you can't have, you know, such and such facial hair, no beards, you know, and just really detailed. A: A: I don't know that that would be a good environment to work in. *Hypothesis:* that would be a good environment to work in Heuristics: C V. BERT: C J. BERT: D AAs: C {C, C, C} Gold: D [2, 0, 0, 0, 0, -1, -2, -3]
7	*Premise:* "Willy did mention it. I was puzzled, I 'll admit, but now I understand." How did you know Heather had been there? *Hypothesis:* Heather had been there Heuristics: N V. BERT: E J. BERT: E AAs: E {E, E, E} Gold: D [3, 3, 3, 2, 1, 1, 0, 0, 0]

Table 4: Models' predictions for CB test items. Labels in { } are predictions by individual AAs. E: entailment, C: contradiction, N: neutral, D: disagreement.

5 Analysis

Table 3 also gives F1 for each class on the test set. AAs outperform all BERT-based models under all classes. However, compared with the Heuristic, AAs show an inferior result on "Neutral" items mainly due to the lack of "Neutral" training data. The first 4 examples in Table 4 show examples for which AAs make the correct prediction while other baselines might not. The confusion matrix in Table 5 shows that the majority (∼60%) of errors come from wrongly predicting a systematic inference item as a disagreement item. In 91% of

Predict \ Gold	E	N	C	D	Total
E	37	2	0	13	52
N	1	10	0	3	14
C	0	0	34	13	47
D	20	7	20	80	127
Total	58	19	54	109	240

Table 5: Confusion matrix for the test set. E: entailment, N: neutral, C: contradiction, D: disagreement.

	negation	modal	conditional	question	negR
Heuristic	51.29	48.02	37.69	44.64	54.16
V. BERT	60.91	73.98	44.84	53.02	61.91
J. BERT	60.94	73.95	46.02	51.68	63.67
AAs	**65.96**	**80.18**	**48.05**	**54.95**	**68.00**

Table 6: F1 for CB test set under the embedding environments and "I don't know/believe/think" ("negR").

such errors, AAs predict that there is more than one mode for the annotation (i.e., the three labels predicted by individual "annotators" in AAs are not unanimous), as in example 5 in Table 4. AAs are thus predicting more modes than necessary when the annotation is actually following a uni-modal distribution. On the contrary, when the item is supposed to be a disagreement item but is missed by AAs (as in example 6 and 7 in Table 4), AAs mistakenly predict that there is only one mode in the annotations 78% of the time. It thus seems that a method which captures accurately the number of modes in the annotation distribution would lead to a better model.

We also examine the model performance for different linguistic constructions to investigate whether the model learns some of the linguistic patterns present in the Heuristic baseline. The Heuristic rules are strongly tied to the embedding environments. Another construction used is one which can lead to "neg-raising" reading, where a negation in the matrix clause is interpreted as negating the content of the complement, as in example 3 (Table 4) where *I* ***do not think*** *they have seen a really high improvement* is interpreted as *I think they* ***did not see*** *a really high improvement*. "Neg-raising" readings often occur with *know*, *believe* or *think* in the first person under negation. There are 85 such items in the test set: 41 contradictions (thus neg-raising items), 39 disagreements and 5 entailments. Context determines whether a neg-raising inference is triggered (An and White, 2019).

Correct inference by Heuristic?	Yes (130)		No (110)	
	Acc.	F1	Acc.	F1
V. BERT	80.00	80.45	41.51	42.48
J. BERT	79.74	80.04	42.73	44.15
AAs	**84.37**	**84.85**	**46.97**	**48.75**

Table 7: BERT-based models performance on test items correctly predicted by vs. items missed by linguistic rules. Numbers next to Yes/No denote the size.

Table 6 gives F1 scores for the Heuristic, BERT models and AAs for items under the different embedding environments and potential neg-raising items in the test set. Though AAs achieve the best overall results, it suffers under conditional and question environments, as the corresponding training data is scarce (9.04% and 14.17%, respectively). The Heuristic baseline always assigns contradiction to the "I don't know/believe/think" items, thus capturing all 41 neg-raising items but missing disagreements and entailments. BERT, a SOTA NLP model, is not great at capturing such items either: 71.64 F1 on contradiction vs. 52.84 on the others (Vanilla BERT); 71.69 F1 vs. 56.16 (Joint BERT). Our AAs capture neg-raising items better with 77.26 F1 vs. 59.38, showing an ability to carry out context-dependent inference on top of the learned linguistic patterns. Table 7, comparing performance on test items correctly predicted by the linguistic rules vs. items for which context-dependent reasoning is necessary, confirms this: AAs outperform the BERT baselines in both categories.

6 Conclusion

We introduced finer-grained natural language inference. This task aims at teasing systematic inferences from inherent disagreements, overlooked in prior work. We show that our proposed AAs, which simulate the uncertainty in annotation process by capturing the modes in annotations, perform statistically significantly better than all baselines. However the best performance obtained (∼66%) is still far from achieving robust NLU, leaving room for improvement.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This material is based upon work supported by the National Science Foundation under Grant No. IIS-1845122.

References

Hannah Youngeun An and Aaron Steven White. 2019. The lexical and grammatical sources of neg-raising inferences. In *Proceedings of the Society for Computation in Linguistics (SCiL 2020)*, pages 220–233.

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, EMNLP 2015*, pages 632–642.

Adrian Calma and Bernhard Sick. 2017. Simulation of annotators for active learning: Uncertain Oracles. In *Proceedings of the Workshop and Tutorial on Interactive Adaptive Learning co-located with European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2017)*, pages 49–58.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005*, pages 177–190.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. In *Sinn und Bedeutung 23*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 4171–4186.

William Gantt, Benjamin Kane, and Aaron Steven White. 2020. Natural language inference with mixed effects. In *The Ninth Joint Conference on Lexical and Computational Semantics (*SEM 2020)*.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019a. Do you know that Florence is packed with visitors? Evaluating state-of-the-art models of speaker commitment. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 4208–4213.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019b. 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, EMNLP 2019*, pages 6085–6090.

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, NAACL-HLT 2018*, pages 1886–1895.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*.

Bill MacCartney and Christopher D. Manning. 2009. An extended model of natural logic. In *Proceedings of the Eight International Conference on Computational Semantics, IWCS 2009*, pages 140–156.

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 Conference of the Association for Computational Linguistics, ACL 2019*, pages 3428–3448.

Ellie Pavlick and Chris Callison-Burch. 2016. Most "babies" are "little" and most "problem" are "huge": Compositional entailment in adjective-nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, pages 677–694.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1532–1543.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? Adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2019*, pages 7–14.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

H. Sebastian Seung, Manfred Opper, and Haim Sompolinsky. 1992. Query by committee. In *Proceedings of the Fifth Annual ACM Conference on Computational Learning Theory, COLT 1992, Pittsburgh, PA, USA, July 27-29, 1992*, pages 287–294.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference*

on Neural Information Processing Systems 2019, NeurIPS 2019, pages 3261–3275.

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, NAACL-HLT 2018*, pages 1112–1122.

Appendix A Linguistic Rules

Our linguistic rules are inspired by and adapted from Jiang and de Marneffe (2019b) to explicitly include the most discriminating expressions for disagreement items. We utilize three linguistic features which are provided in CB: entailment-canceling environment (negation, modal, question, antecedent of conditional), matrix verb and its subject person.

1. Items under conditional are disagreement.
2. Items under question and with second person are neutral.
3. Items under question and with non-second person are disagreement.
4. Items of the form "I don't know/think/believe" are contradiction (i.e., negRaising structure).
5. Items with factive verbs are entailment.
6. Items under negation and with non-factive verbs are disagreement.
7. Items under modal and with non-third person are entailment.

When this policy is executed, there are two additional auxiliary rules: Items not falling in any group above are assigned a disagreement label as it is the dominant class in CB; For items satisfying more than one rule, the label will be determined by the higher-ranked rule (i.e., a smaller number indicates a higher rank). Note that the rules above also reveal the most discriminating expressions for each class.

Appendix B Impact of r

We experiment with three different r values, 0.25, 0.4, 0.7. Intuitively,

- 0.25: each module contributes equally;
- 0.4: finer-grained NLI predictor is the main component and three artificial annotators are to complement the main predictor;
- 0.7: over-amplify the role of main predictor and suppress three artificial annotators.

Table A1 shows the results on the dev set in terms of both accuracy and F1 under different r values. We set r to 0.4 as it achieves the best accuracy on the dev set.

r	0.25	0.4	0.7
Accuracy	65.33	**65.67**	65.33
F1	64.86	64.57	**65.06**

Table A1: Performance of our proposed AAs on the dev set under different r values.

	Train	Test	Overall
CBOW	35	30	65
Vanilla BERT	205	7	212
Ensemble BERT	645	10	655
AAs (`ours`)	955	9	964

Table A2: The training (10 epochs) and testing time as well as overall running time for each neural network-based model. All numbers are in seconds.

Appendix C Reproducibility Checklist

C.1 Implementation Details

We set r in eq. 2 at 0.4 (see Appendix B). For all experiments, we use the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 1e-5 and fine-tune up to 10 epochs. The batch size of first three baselines is 8 and that of Ensemble BERT baseline and our AAs is 2. The gradient is clipped when its norm exceeds 5. We select the best model for each method using the accuracy on the dev set, and the average performance on dev set is shown in Table 3 as well. Hyperparameters for underlying BERT (`bert-base-uncased`)[5] are the default.

C.2 Summary Statistics of Results

For the results reported in Table 3, we run each baseline (except for "Always 0" and Heuristic rules) and AAs 10 times, and report the average. For the results reported in Table 6 and 7, we randomly select 3 trained models for each baseline and AAs, and report the average. We also share the checkpoints of AAs at `https://github.com/FrederickXZhang/FgNLI` to help reproduce the results given in Section 5.

C.3 Computational Resources

All experiments are conducted using one single GeForce GTX 2080 Ti 12 GB GPU (with significant CPU resources). The overall running time of

[5] `https://huggingface.co/transformers/`

baselines and AAs are listed in Table A2.

C.4 Dataset

The characteristics of the CommitmentBank (CB) are detailed in Section 2. The original version is available at https://github.com/mcdm/CommitmentBank, and the recast version used in this work is available at https://github.com/FrederickXZhang/FgNLI.