Inherent Disagreements in Human Textual Inferences

Ellie Pavlick
Brown University
ellie_pavlick@brown.edu

Tom Kwiatkowski
Google Research
tomkwiat@google.com

Abstract

We analyze human's disagreements about the validity of natural language inferences. We show that, very often, disagreements are not dismissible as annotation ''noise'', but rather persist as we collect more ratings and as we vary the amount of context provided to raters. We further show that the type of uncertainty captured by current state-of-the-art models for natural language inference is not reflective of the type of uncertainty present in human disagreements. We discuss implications of our results in relation to the recognizing textual entailment (RTE)/natural language inference (NLI) task. We argue for a refined evaluation objective that requires models to explicitly capture the full distribution of plausible human judgments.

1 Introduction

Entailment is arguably one of the most fundamental of language understanding tasks, with Montague himself calling entailment ''the basic aim of semantics'' (Montague, 1970). Computational work on recognizing textual entailment (RTE) (also called natural language inference, or NLI) has a long history, ranging from early efforts to model logical phenomena (Cooper et al., 1996), to later statistical methods for modeling practical inferences needed for applications like information retrieval and extraction (Dagan et al., 2006), to current work on learning common sense human inferences from hundreds of thousands of examples (Bowman et al., 2015; Williams et al., 2018).

Broadly speaking, the goal of the NLI task is to train models to make the inferences that a human would make. Currently, ''the inferences that a human would make'' are determined by asking multiple human raters to label pairs of sentences, and then seeking some consensus among them. For example, having raters choose among discrete labels and taking a majority vote (Dagan et al., 2006; Bowman et al., 2015; Williams et al., 2018), or having raters use a continuous Likert scale and taking an average (Pavlick and Callison-Burch, 2016a; Zhang et al., 2017). That is, the prevailing assumption across annotation methods is that there is a single ''true'' inference about h given p that we should train models to predict, and that this label can be approximated by aggregating multiple (possibly noisy) human ratings as is typical in many other labelling tasks (Snow et al., 2008; Callison-Burch and Dredze, 2010).

Often, however, we observe large disagreements among humans about whether or not h can be inferred from p (see Figure 1). The goal of this study is to establish whether such disagreements can safely be attributed to ''noise'' in the annotation process (resolvable via aggregation), or rather are a reproducible signal and thus should be treated as part of the NLI label assigned to the p/h pair. Specifically, our primary contributions are:

- We perform a large-scale study of humans' sentence-level inferences and measure the degree to which observed disagreements persist across samples of annotators.
- We show that current state-of-the-art NLI systems do not capture this disagreement by default (by virtue of treating NLI as probabilistic) and argue that NLI evaluation should explicitly incentivize models to predict distributions over human judgments.
- We discuss our results with respect to the definition of the NLI task, and its increased usage as a diagnostic task for evaluating ''general purpose'' representations of natural language.

Transactions of the Association for Computational Linguistics, vol. 7, pp. 677–694, 2019. https://doi.org/10.1162/tacl_a_00293
Action Editor: Christopher Potts. Submission batch: 5/2019; Published: 11/2019.

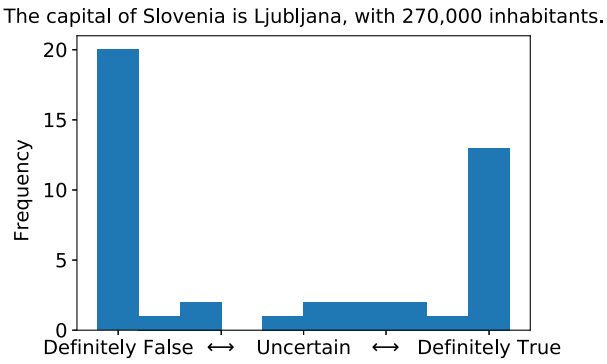

Figure 1: Example p/h pair on which humans exhibit strong disagreements about whether h can be inferred from p. Here, the disagreement appears to stem from the implicature, but we observe similar disagreements on a variety of linguistic phenomena.

2 The RTE/NLI Task

The task of RTE/NLI is fundamentally concerned with drawing conclusions about the world on the basis of limited information, but specifically in the setting when both the information and the conclusions are expressed in natural language. That is, given a proposition p, should one infer some other proposition h to be true?

Traditionally, in formal linguistics, the definition of *entailment* used is that defined in formal logic—namely, p entails h if h is true in every possible world in which p is true. This logical definition takes for granted that lexical and constructional meanings are fixed in such a way that it is possible to fully pre-specify and then repeatedly apply those meanings across all contexts. From the point of view of evaluating NLP systems' ability to reason about entailment, these are clearly difficult criteria to operationalize. Thus, within NLP, we have rarely if ever evaluated directly against this definition. Rather, work has been based on the below informal definition:

> p entails h if, typically, a human reading p would infer that h is most likely true. . . [assuming] common human understanding of language [and] common background knowledge (Dagan et al., 2006).

This definition was intended to undergo refinement overtime, with Dagan et al. (2006) explicitly stating that the definition was ''clearly not mature yet'' and should evolve in response to observed shortcomings, and, in fact, substantial discussion surrounded the original definition of the RTE task. In particular, Zaenen et al. (2005) argued that the definition needed to be made more precise, so as to circumscribe the extent to which ''world knowledge'' should be allowed to factor into inferences, and to explicitly differentiate between distinct forms of textual inference (e.g., entailment vs. conventional implicature vs. conversational implicature). Manning (2006) made a counter-argument, pushing back against a prescriptivist definition of what types of inferences are or are not licensed in a specific context, instead advocating that annotation tasks should be ''natural'' for untrained annotators, and that the role of NLP should be to model the inferences that humans make in practical settings (which include not just entailment, but also pragmatic inferences such as implicatures). Both supported the use of the term ''inference'' over ''entailment'' to acknowledge the divergence between the working NLP task definition and the notion of entailment as used in formal semantics.[1]

Since the task's introduction, there has been no formal consensus around which of the two approaches offers the better cost–benefit tradeoff: precise (at risk of being impractical), or organic (at risk of being ill-defined). That said, there has been a clear gravitation toward the latter, apparent in the widespread adoption of inference datasets that explicitly prioritize natural inferences over rigorous annotation guidelines (Bowman et al., 2015; Williams et al., 2018), and in the overall shift to the word ''inference'' over ''entailment.'' There has also been significant empirical evidence supporting the argument that humans' semantic inferences are uncertain and context-sensitive (Poesio and Artstein, 2005; Versley, 2008; Simons et al., 2010; Recasens et al., 2011; de Marneffe et al., 2012; Passonneau et al., 2012; Pavlick and Callison-Burch, 2016a,b; Tonhauser et al., 2018, among others) suggesting computational models would benefit from focusing on ''speaker meaning'' over ''sentence meaning'' when it comes to NLI (Manning, 2006; Westera and Boleda, 2019).

Thus, in this paper, we assume that NLP will maintain this hands-off approach to NLI, avoiding definitions of what inferences humans *should* make or which types of knowledge they *should* invoke. We take the position that, ultimately, our

[1] We, too, adopt the word ''inference'' for this reason.

goal in NLP is to train models that reverse-engineer the inferences a human would make when hearing or reading language in the course of their daily lives, however ad-hoc the process that generates those inferences might be. Therefore, our question in this paper is not yet *what process* humans use to draw inferences from natural language, but merely: Left to their own devices, do humans, in general, tend to follow *the same process*? Note that this question is independent of the decision of whether to treat annotations as discrete versus gradable. Even if NLI is treated as a gradable phenomenon (as we believe it should be), a world in which all humans share the same notion of uncertainty necessitates very different models, annotation practices, and modes of evaluation than a world in which people may disagree substantially in specific situations, use different heuristics, and/or have different preferences about how to resolve uncertainty. Specifically, current practices—in which we aggregate human judgments through majority vote/averaging and evaluate models on their ability to predict this aggregated label—are only appropriate if humans all tend to use the same process for resolving uncertainties in practice.

3 NLI Data and Annotation

To perform our analysis, we collect NLI judgments at $50\times$ redundancy for sentence pairs drawn from a variety of existing NLI datasets. Our annotation procedure is described in detail in this section. All of the data and collected annotations are available at `https://github.com/epavlick/NLI-variation-data`.

3.1 Sentence Pairs

We draw our p/h pairs from the training sets of each of the following five datasets: RTE2 (Dagan et al., 2006), SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), JOCI (Zhang et al., 2017), and DNC (Poliak et al., 2018b). Table 1 shows randomly sampled positive ($p \rightarrow h$) and negative ($p \not\rightarrow h$) examples from each. These datasets differ substantially in the procedures used to generate the data, and in the types of inferences they attempt to test. RTE2 consists of premises/hypothesis pairs derived predominantly from the output of information retrieval systems run over newswire text and annotated by experts (researchers in the field). SNLI consists of premises derived from image captions with hypotheses written and judged by non-expert (crowdsourced) annotators. MNLI was constructed in the same way as SNLI but contains premises drawn from a range of text genres, including letters, fiction, and telephone conversations. JOCI is intended to target ''common sense'' inferences, and contains premises drawn from existing NLI datasets[2] paired with hypothesis that were automatically generated via either templates or seq2seq models and then refined by humans. The DNC consists predominantly of naturally occurring premises paired with template-generated hypotheses, and comprises a number of sub-corpora aimed at testing systems' understanding of specific linguistic phenomena (e.g., lexical semantics, factuality, named entity recognition). We draw from this variety of datasets in order to ensure a diversity of types of textual inference and to mitigate the risk that the disagreements we observe are driven by a specific linguistic phenomenon or dataset artifact on which humans' interpretations particularly differ.

We sample 100 p/h pairs from each dataset. In every dataset, we limit to pairs in which the premise and the hypothesis are both less than or equal to 20 words, to minimize cognitive load during annotation. We attempt to stratify across expected labels to ensure an interesting balance of inference types. For RTE2, SNLI, and MNLI, this means stratifying across three categories (ENTAILMENT/CONTRADICTION/NEUTRAL). For JOCI, the p/h pairs are labeled on a five-point Likert scale, where 1 denotes that h is ''impossible'' given p and 5 denotes that h is ''very likely'' given p, and thus we stratify across these five classes. In the DNC, all sub-corpora consist of binary labels (ENTAILMENT/NON-ENTAILMENT) but some sub-corpora contain finer-grained labels than others (e.g., three-way or five-way labels). Thus, when sampling, we first stratify across sub-corpora[3] and then across the most fine-grained label type available for the given sub-corpus.

3.2 Annotation

We show each p/h pair to 50 independent raters on Amazon Mechanical Turk. We ask them to

[2]We skip the subset of JOCI that was drawn from SNLI, to avoid redundancy with our own SNLI sample.

[3]We skip two sub-corpora (VerbCorner and Puns), the former because it contains nonced words and thus is difficult to ask humans to label without some training, and the latter because of the potential for noisy labels due to the fact that some people, bless their hearts, just don't appreciate puns.

SNLI	Three dogs on a sidewalk. → There are more than one dog here. A red rally car taking a slippery turn in a race. → ¬ The car is stopped at a traffic light.
MNLI	Historical heritage is very much the theme at Ichidani. → Ichidani's historical heritage is important. okay i uh i have five children all together → ¬ I do not have any children.
RTE2	Self-sufficiency has been turned into a formal public awareness campaign in San Francisco, by Mayor Gavin Newsom. → Gavin Newsom is a politician of San Francisco. The unconfirmed case concerns a rabies-like virus known only in bats → ¬ A case of rabies was confirmed.
JOCI	It was Charlie 's first day of work at the new firm → The firm is a business. A young girl is holding her teddy bear while riding a pony . → ¬ The bear attacks.
DNC	Tony bent the rod. → Tony caused the bending. When asked about the restaurant, Jonah said, 'Sauce was tasteless.' ↛ Jonah liked the restaurant.

Table 1: Examples of p/h pairs from each of our source datasets. The top pair is one labeled by the original dataset as a valid inference (one that should be drawn), the bottom as an invalid inference (either h is contradictory given p ($p \rightarrow \neg h$), or h simply cannot be inferred ($p \not\rightarrow h$)). For DNC, examples shown are from the VerbNet (top) and Sentiment (bottom) sub corpora.

indicate using a sliding bar, which ranges from -50 to 50,[4] how likely it is that h is true given that p is true, where -50 means that h is definitely not true ($p \rightarrow \neg h$), 50 means that h is definitely true ($p \rightarrow h$), and 0 means that h is consistent with but not necessarily true given p ($p \not\rightarrow h$). Raters also have the option to indicate with a checkbox that either/both of the sentences do not make sense and thus no judgment can be made. We attempt to pitch the task intuitively and keep the instructions light, for reasons discussed in Section 2. We provide brief instructions followed by a few examples to situate the task. Our exact instructions and examples are shown Table 2.

Raters label pairs in batches of 20, meaning we have a minimum of 20 ratings per rater. We pay $0.30 per set of 20. We restrict to raters who have a 98% or better approval rating with at least 100 HITs approved, and who are located in a country in which English is the native language (US, Canada, UK, Australia, New Zealand).

For each pair of sentences, **assume that the first sentence (S1) is true, describes a real scenario, or expresses an opinion**. Using your best judgment, **indicate how likely it is that the second sentence (S2) is also true, describes the same scenario, or expresses the same opinion**. If either sentence is not interpretable, check the ''Does Not Make Sense'' box. Several examples are given below.
Example 1: In the below example, the slider is far to the right because we can be very confident that if a person is ''on a beach'' than that person is ''outside''. S1: A woman is on a beach with her feet in the water. S2: The woman is outside.
Example 2: In the below example, the slider is far to the left because we can be very confident that if a person is ''on a beach'' then that person is NOT ''in her living room''. S1: A woman is on a beach with her feet in the water. S2: The woman is in her living room.
Example 3: In the below example, the slider is in the center because knowing that woman is on the beach does not give us any information about the color of her hair and so we cannot reasonably make a judgment about whether or not her hair is brown. S1: A woman is on a beach with her feet in the water. S2: The woman has brown hair.

Table 2: Instructions and examples shown to raters. Raters indicated their responses using a sliding bar which ranged from -50 to 50. In the instructions actually shown, the examples were shown alongside a sliding bar reflecting the desired rating. Exact UI not shown for compactness.

3.3 Preprocessing

Filtering. In total, we had 509 workers complete our HITs, with an average of 2.5 tasks (50 sentence pairs) per worker. We follow the methods from White et al. (2018) and remove workers who demonstrate consistently low correlations with others' judgments. Specifically, for each sentence pair s, for each worker w_i, we compute the Spearman correlation between w_i's labels and every other w_j who labeled s. Across all pairs of workers, the mean correlation is 0.48. We consider a pair of workers on a given assignment to be an outlier if the correlation between those workers' ratings falls outside 1.5 times the interquartile

[4] Raters do not see specific numbers on the slider.

range of all the correlations (White et al., 2018). We find 234 pairs to be outliers, and that they can be attributed to 14 individual workers. We therefore remove all annotations from these 14 workers from our analysis. Additionally, we remove ratings from 37 workers[5] who have fewer than 15 useable data points (i.e., judgments not including cases in which they choose the ''does not make sense'' option), as this will prevent us from properly estimating and thus correcting for their individual annotation bias (described in the following section). Finally, we remove p/h pairs that, after removing all problematic workers and ''does not make sense'' judgments, are left with fewer than 15 judgments. In the end, we have 496 p/h pairs with a mean of 39 labels per pair.

Normalization. One confound that results from collecting annotations on a continuous scale is that each rater may choose to use the scale differently. Thus, we apply z-score normalization to each worker's labels for each assignment, meaning each worker's ratings are rescaled such that the mean across all labels from a single worker within a single batch is 0 and the standard deviation is 1. This normalization is not perfect, as every batch has a slightly different set of pairs, and so normalized scores are not comparable across batches. For example, if, by chance, a batch were to contain mostly pairs for which the ''true'' label was $p \rightarrow h$, a score of zero would imply $p \rightarrow h$, whereas if a batch were to include mostly pairs for with the ''true'' label was $p \rightarrow \neg h$, zero would correspond to $p \rightarrow \neg h$. However, for the purposes of our analysis, this is not problematic; because our interest is comparing disagreements between annotations on each specific p/h pair, it is only important that two worker's labels on the same pair are comparable, not that judgments across pairs are comparable.[6]

[5]Results presented throughout are based on data with these workers removed. However, rerunning analysis with these workers included did not affect our overall takeaways.

[6]On our own manual inspection, it is nearly always the case that the mean (0) is roughly interpretable as neutral, with only moderate deviations from one example to the next. Nonetheless, when interpreting the figures in the following sections, note that the center of one pair's distribution is not necessarily comparable to the center of another's.

4 Analysis of Human Judgments

4.1 Experimental Design

We aim to establish whether the disagreements observed between humans' NLI judgments can be attributed to ''noise'' in the annotation process. We make the assumption that, if the disagreements are attributable to noise, then the observed human judgments can be modeled as a simple Gaussian distribution, where the mean is the true label. This model can account for the fact that some cases might be inherently harder than others—this could, for example, be reflected by higher variance—but, overall, the labels are nonetheless in accordance with the assumption that there exists a fundamentally ''true'' label for each p/h pair which we can faithfully represent via a single label or value, obtainable via aggregation.

For each sentence pair, we randomly split the collected human labels into train and test. Specifically, we hold out 10 labels from each pair to use as our test set. The training data are composed of the remaining labels, which varies in number from 5 to 40, depending on how many labels were left for that pair after preprocessing (see Section 3.3). The average number of training labels is 29. For each sentence pair, we use the training data to fit two models: 1) a single Gaussian and 2) a Gaussian Mixture Model where the number of components is chosen during training,[7] meaning that the model may still choose to fit only one component if appropriate. We compute the log likelihood assigned to the held-out test data under each model, and observe how often, and to what extent, the additional components permitted by the GMM yield a better fit for the held out judgments.

If the mixture model frequently choses to use more than one effective component, and if doing so results in a better fit for the held-out data than the unimodal Gaussian, we interpret this as evidence that, for many sentence pairs, human judgments exhibit reproducibly multimodal distributions. Thus, for such sentence pairs, the current practice of aggregating human judgments into a single label would fail to accurately capture the types

[7]We use the Variational Bayesian estimation of a Gaussian mixture provided in SciKit learn, with the maximum number of components set to be the number of points in the training data: `https://scikit-learn.org/stable/modules/generated/sklearn.mixture.BayesianGaussianMixture.html`

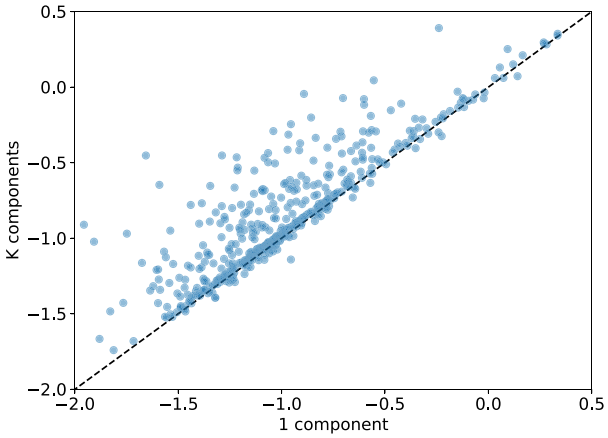

Figure 2: Log likelihood assigned to test data under the single-component Gaussian (x-axis) vs the k-component GMM (y-axis). Results show an average over 10 random train/test splits; error bars not shown to reduce clutter. Overall, multimodal distributions generalize better to unseen human judgments than do single Gaussians.

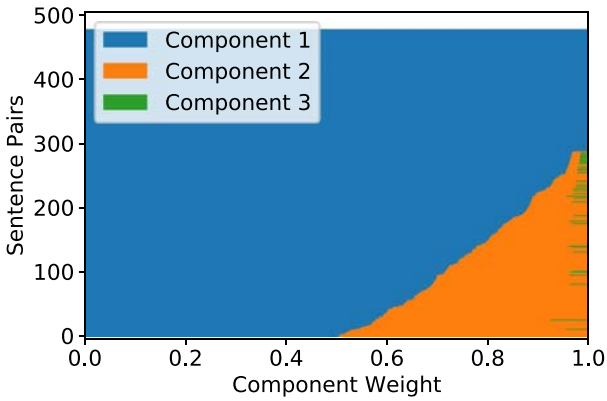

Figure 3: Weights of effective components for each p/h pair. y-axis corresponds to the pairs in our data, sorted by weight of the second component. The figure should be interpreted as follows: When the line is all blue (pair #400), the GMM found a single component with a weight of 1. When the line contains mixed colors, the model found multiple components with the depicted weights (e.g., pair #0 has two components of equal weight).

of semantic inferences that humans might make about the given p/h pair.

4.2 Results

Are distributions unimodal? Figure 2 shows, for each sentence pair, the test log likelihood under the one-component Gaussian model versus the k-component GMM. If the data were in fact sampled from an underlying distribution defined by a single Gaussian, we would expect the points to be distributed approximately randomly around the $y = x$ line. That is, most of the time the GMM would provide no advantage over the single Gaussian. What we see instead is that the majority of points fall on or above the $y = x$ line, indicating that, when there is a difference, the additional components deemed necessary in training tend to generalize to unseen human judgments. Very few points fall below the $y = x$ line, indicating that when models choose to fit multiple components, they are correctly modeling the true data distribution, rather than overfitting the training set. We note that the majority of points fall on $y = x$, indicating that most examples *do* exhibit consensus around one ‘‘true’’ label.[8] Figure 3 shows, for each sentence pair, the weights of the effective components according the the GMM. We see that for 20% of the sentence pairs, there is a nontrivial second component (weight > 0.2), but rarely are there more than two components with significant weights.

Figure 4 shows several examples of sentences for which the annotations exhibit clear bimodal distributions. These examples show the range of linguistic phenomena[9] that can give rise to uncertainty. In the first example, from SNLI, there appears to be disagreement about the degree to which two different descriptions could potentially refer to the same scenario. In the second example, from DNC and derived from VerbNet (Chklovski and Pantel, 2004), there is disagreement about the manner aspect of *‘‘swat’’*, that is, whether or not *‘‘swatting’’* is necessarily *‘‘forceful’’*. In the third example, from DNC and derived from the MegaVerdicality dataset (White and Rawlins, 2017), there appears to be disagreement about the degree to which *‘‘confess that’’* should be treated as factive.

These examples highlight legitimate disagreements in semantic interpretations, which can be difficult to control without taking a highly prescriptivist approach to annotation. Doing so,

[8]We verified that, if forced to fit more than one component, the model often overfits, confirming that these examples are indeed best modeled as unimodal distributions.

[9]By corpus, RTE exhibits the least variation and JOCI exhibits the most, though all of the corpora are comparable. We did not see particularly interesting trends when we broke down the analysis by corpus explicitly, so, for brevity, we omit the finer-grained analysis.

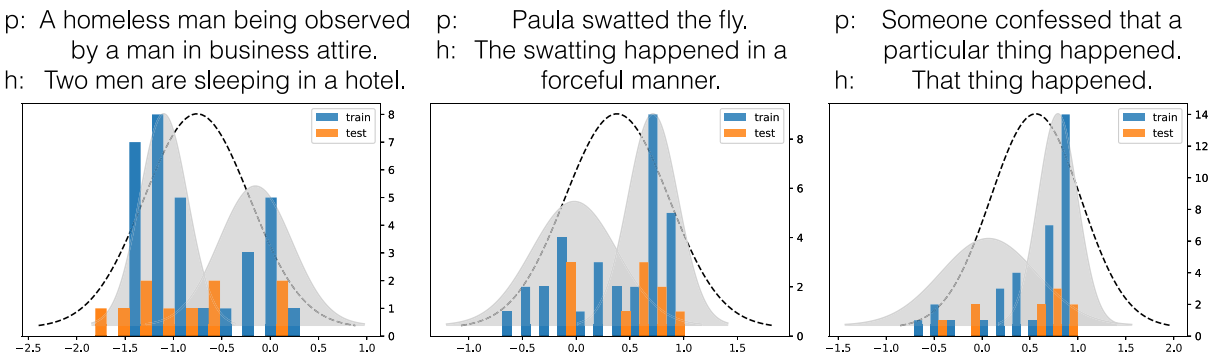

Figure 4: Examples of sentence pairs with bi-modal human judgment distributions. Examples are drawn from SNLI, the VerbNet portion of DNC, and the MegaVerdicality portion of DNC (from left to right). Training distribution is in blue; test in orange. Dotted black line shows the model fit when using a single component; shaded gray shows the model learned when allowed to fit k components. Distributions are over z-normalized scores in which 0 roughly corresponds to neutral ($p \not\rightarrow h$) but not precisely (§3.3).

however, would compromise both the ‘‘naturalness’’ of the task for annotators and the empiricist approach to representation learning currently desired in NLP (as discussed in Section 2).

Does context reduce disagreement? One fair objection to these results is that sentence-level inferences are problematic due to the lack of context provided. It is reasonable to believe that the divergences in judgments stem from the fact that, when details of the context are left unspecified, different raters choose to fill in these details differently. This would inevitably lead to different inferences, but would not be reflective of differences in humans' representations of linguistic ‘‘meaning’’ as it pertains to NLI. We thus explore whether providing additional context will yield less-divergent human judgments. To do this, we construct a small dataset in which we can collect annotations with varying levels of context, as described next.

Method. We sample sentences from Wikipedia, restricting to sentences that are at least four words long and contain a subject and a verb. We consider each of these sentences to be a candidate premise (p), and generate a corresponding hypothesis (h) by replacing a word w_1 from p with a substitute w_2, where w_2 has a known lexical semantic relationship to w_1. Specifically, we use as set of 300 word pairs: 100 hypernym/hyponym pairs, 100 antonym pairs, and 100 co-hyponym pairs. We chose these categories in order to ensure that our analysis consists of meaningful substitutions and that it covers a variety of types of inference judgments. Our hypernyms and antonyms are taken from WordNet (Fellbaum, 1998), with hypernyms limited to first-sense immediate hypernyms. Our co-hyponyms are taken from an internal database, which we constructed by running Hearst patterns (Hearst, 1992) over a large text corpus. The 300 word pairs we used are available for inspection at https://github.com/epavlick/NLI-variation-data. After making the substitution, we score each candidate p and h with a language model (Józefowicz et al., 2016) and disregard pairs for which the perplexity of h is more than 5 points above that of p. This threshold was chosen based on manual inspection of a sample of the output, and is effective at removing sentences in which the substitution yielded a meaningless hypothesis—for example, by replacing a w_1 that was part of a multiword expression.

For each resulting p/h pair, we collect ratings at three levels: word level, in which p and h are each a single word; sentence level, in which p and h are each a sentence; and paragraph level, in which p is a full paragraph and h is a sentence (as depicted in Figure 6). We use the same annotation design as described in Section 3.2. To quantify the level of disagreement in the observed judgments, we compute two measures: 1) variance of observed ratings and 2) Δ log likelihood, that is, the change in log likelihood of held out data that results from using a k-component GMM over a single-component Gaussian (as described in the previous section). We note that Δ log likelihood is a more direct measure of the type of disagreement in which we are interested in this paper (i.e., disagreements stemming from multimodal distributions of judgments that are

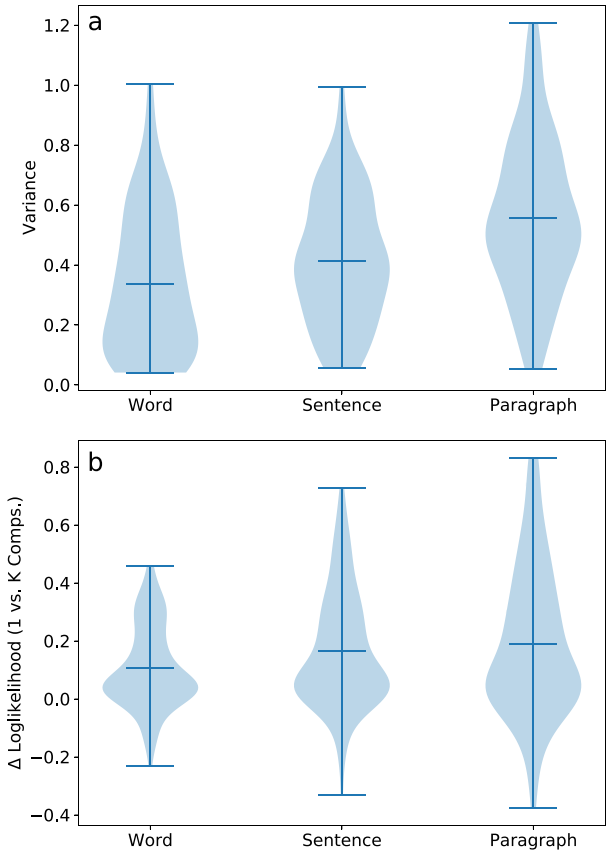

Figure 5: Distributions of variances (top) and Δ log likelihood (bottom) for human ratings resulting from word, sentence, and paragraph contexts. The average variances of all levels are significantly different at $p < 0.05$ (word < sentence < paragraph). Average ΔLL for words was significantly lower than for sentences and paragraphs, but there is no significant difference between sentences and paragraphs.

not well summarized by a single label/value). High variance distributions may correspond to ‘‘difficult’’ cases which are nonetheless still unimodal.

Results. Figure 5 shows the distribution of each metric as a function of the level of context given to raters. The trend is counter to our initial intuition: Both measures of disagreement actually increase when raters see more context. On average, we see a variance of 0.34 ± 0.02 when raters are shown only words, 0.41 ± 0.02 when raters are shown sentences, and 0.56 ± 0.02 when raters are given a full paragraph of context (95% confidence intervals). The trend for Δ log likelihood is similar: Disagreement at the word level (0.11 ± 0.02) is significantly lower than at the sentence (0.21 ± 0.04) and paragraph (0.22 ± 0.03) level, though there is no significant difference in Δ log likelihood between sentence-level and paragraph-level.

Figure 6 shows an example p/h pair for which additional context increased the variance among annotators. In the example shown, humans are generally in agreement that *‘‘boating’’* may or may not imply *‘‘picknicking’’*, when no additional context is given. However, when information is provided which focuses on boating on a specific canal, emphasizing the activities that the water itself is used for, people diverge in their inference judgments, with one group centered around contradiction and a smaller group centered around neutral.

We interpret these results as preliminary evidence that disagreement is not necessarily controllable by providing additional context surrounding the annotation (i.e., we do not see evidence that increasing context helps, and it may in fact hurt). We hypothesize that, in fact, less context may result in higher agreement due to the fact that humans can more readily call on conventionalized ‘‘default’’ interpretations. For example, in the case of single words, people likely default

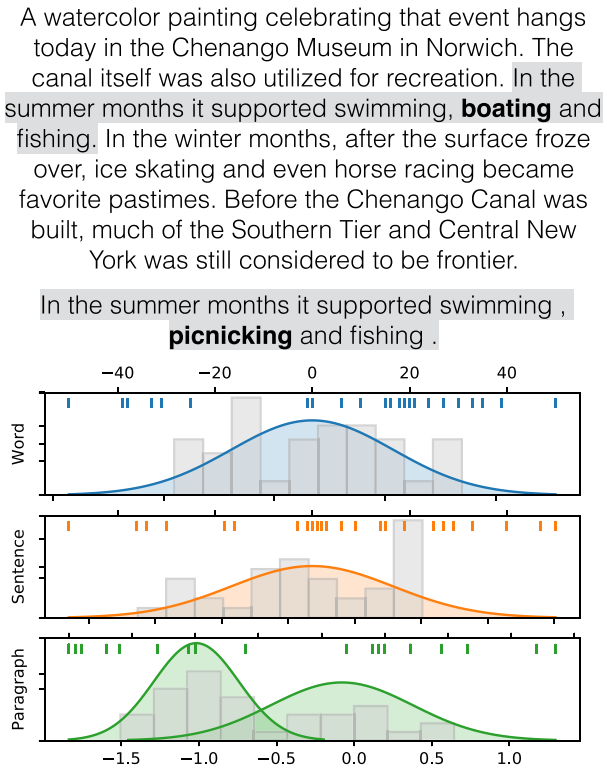

Figure 6: In the word case, human judges were shown only the words (bolded); in the sentence case, judges were shown pairs of sentences (gray highlight); in the paragraph case, judges were shown all of the text. Judges did not see markup (bold/highlight) when presented the text to judge. Gray bars show distribution of z-normalized scores, ticks show raw (unnormalized) scores, bell curves are estimated by the GMM.

to reading them as referring expressions for a single entity/event, and thus make judgments consistent with the prototypical lexical entailment relations between these words. Additional context provides increased opportunity for inferences based on pragmatics and world knowledge (e.g., inferences about the question under discussion and the speaker's intent), which are less likely to follow consistent conventions across all raters.

We consider this study exploratory, as there are some confounds. Most notably, increasing the amount of context clearly increases cognitive load on annotators, and thus we would expect to see increased variance even if there were no increase in actual interpretive disagreements. However, the increase in the Δ log likelihood metric is more meaningful, because randomly distributed noise (which we might expect in the case of high cognitive load/low annotator attention) should lead to higher variance but not multimodality. More work is needed to explore this trend further, and to determine whether increasing context would be a viable and productive means for reducing disagreements on this task.

5 Analysis of Model Predictions

5.1 Motivation

Another natural question arising from the analysis presented thus far is whether the phenomenon under investigation even poses a problem for NLP systems at all. That is, whether or not humans' judgments can be summarized by a single aggregate label or value might be a moot question, since state-of-the-art models do not, in practice, predict a single value but rather a distribution over values. It may be the case that these predicted distributions already reflect the distributions observed in the human judgments and thus that the models can be viewed as already adequately capturing the aspects of semantic uncertainty that cause the observed human disagreements. We thus measure the extent to which the softmax distributions produced by a state-of-the-art NLI model trained on the dataset from which the p/h pairs were drawn reflects the same distribution as our observed human judgments.

5.2 Experimental Design

Data. NLI is standardly treated as a classification task. Thus, in order to interface with existing NLI models, we discretize[10] our collected human judgments by mapping the raw (unnormalized) score (which is between -50 and 50) into K evenly sized bins, where K is equal to the number of classes that were used in the original dataset from which the p/h pair was drawn. Specifically, for pairs drawn from datasets which use the three-way ENTAILMENT/CONTRADICTION/NEUTRAL labels (i.e., SNLI, MNLI, and RTE2), we consider human scores less than -16.7 to be CONTRADICTION, those greater than 16.7 to be ENTAILMENT, and those in between to be NEUTRAL. For the binary tasks (DNC), we use the same three-way thresholds, but consider scores below 16.7 to be NONENTAILMENT and those above to be ENTAILMENT. After some experimentation, we ultimately choose to map the

[10]We experimented with multiple variations on this mapping, including using the z-normalized (rather than the raw) human scores, and using bins based on percentiles rather than evenly spaced over the full range. None of these variants noticeably affected the results of our analysis or the conclusions presented in the following section.

	Orig./ Ours	BERT/ Orig	BERT/ Ours	$\cap$
SNLI	0.790	0.890	0.830	76
MNLI	0.707	0.818	0.687	62
RTE2	0.690	0.460	0.470	36
DNC	0.780	0.900	0.800	74
JOCI	0.651	0.698	0.581	41

Table 3: Left to right: Agreement between datasets' original labels and the majority label according to our (discretized) re-annotation; accuracy of BERT NLI model against original labels; accuracy of BERT against re-annotation labels; number of p/h pairs (out of 100) on which all three label sources (original, re-annotation, model prediction) agree on the most likely label. Our analysis in §5.3 is performed only over pairs in $\cap$.

JOCI scores to a three-way classification scheme as well, rather than the original five-way scheme, using $1 =$ CONTRADICTION, $\{2,3,4\} =$ NEUTRAL, and $5 =$ ENTAILMENT. This decision was made after observing that, although our overall results and conclusions remained the same regardless of the way we performed the mapping, the three-way mapping led to higher levels of agreement between the original labels and our newly collected labels, and thus gave the model the best chance of learning the distribution against which it will be tested.[11] Agreement between the original labels (i.e., those in the published version of the data) and our discretized newly collected labels are given in the first column of Table 3. We note that measuring agreement and model accuracy in terms of these discrete distributions is not ideal, and it would be preferable to train the model to directly predict the full distributions, but because we do not have sufficient training data to do this (we only collected full distributions for 100 p/h pairs per dataset) we must work in terms of the discrete labels provided by the existing training datasets.

Model. We use pretrained BERT (Devlin et al., 2019),[12] fine-tuned on the training splits of the datasets from which our test data was drawn. That is, we fine-tune BERT five times, once on each dataset, and then test each model on the subset of our re-annotated p/h pairs that were drawn from the dataset on which it was fine-tuned. We remove from each training set the 100 p/h pairs that we had re-annotated (i.e., the data we use for testing). We use the BERT NLI model off-the-shelf, without any changes to architecture, hyperparameters, or training setup.

Table 3 shows the accuracy of each model on the test set (i.e., our 100 re-annotated sentences) when judged against 1) the original (discrete) label for that pair given in the standard version of the dataset (i.e., the same type of label on which the model was trained) and 2) our new (discretized) label derived from our re-annotation. Table 3 also gives the agreement between the original discrete labels and the discretized re-annotation labels.

Metrics. We want to quantify how well the model's predicted softmax distribution captures the distribution over possible labels we see when we solicit judgments from a large sample of annotators. To do this, we consider the model softmax to be a basic multinomial distribution, and compute 1) the probability of the observed human labels under that multinomial and 2) the cross-entropy between the softmax and the observed human distributions. As a point of comparison, we compute the same metrics for a random sample, of equal size to the set of observed labels, drawn from the multinomial defined by the softmax.

We focus only on p/h pairs on which all three label sources (i.e., the original label provided by the official dataset, the new label we produce by taking the majority vote of our newly collected, discretized human judgments, and the model's prediction) agree. That is, because we want to evaluate whether the model captures the distribution (not just the majority class that it was trained to predict) we want to focus only on cases where it at least gets the majority class right. Because we want to compare against the full distribution of discretized human labels we collected, we don't want to consider cases where the majority class according to this distribution disagrees with the majority class according to the model's training data, since this would unfairly penalize the model. Table 3 shows the number of pairs (out of 100) on which these three label sources agree, for each dataset.

5.3 Results

Overall, the softmax is a poor approximation of the distribution observed across the human

[11] We also try removing JOCI from our analysis entirely, since it is the noisiest dataset, and still reach the same conclusions from our subsequent analysis.

[12] https://github.com/google-research/bert

	Cross Ent.		Log Prob.	
Exp.	0.03	(0.03, 0.03)	−1.6	(−1.7, −1.5)
Obs.	0.37	(0.33, 0.42)	−21.5	(−22.6, −20.1)

Table 4: Softmax is not a good estimate of the distribution of human labels. Exp. refers to the similarity values we expect due to random variation (i.e., what we get when we compute against a random sample drawn from the multinomial defined by the softmax). Obs. refers to the similarity values between the softmax distribution and the human distribution. Numbers in parentheses give 95% confidence intervals. Results are effectively the same for each of individual corpora, so we report only the aggregate results.

judges. The log probability assigned to the observations (i.e., the set of human labels) by the predicted (softmax) multinomial is significantly and substantially lower than the probability that we would expect to be assigned if the observations had been in fact sampled from the predicted distribution. Similarly, the cross entropy between the predicted and the observed distribution is significantly higher than what can be attributed to random noise (Table 4).

Figure 7 shows some examples of p/h pairs for which the softmax substantially misrepresents the nature of the uncertainty that exists among the human labels, in one case because the model predicts with certainty when humans find the judgment ambiguous (due to the need to resolve an ambiguous co-reference) and in the other because the model suggests ambiguity when humans are in clear consensus. Overall, the results indicate that while softmax allows the model to represent uncertainty in the NLI task, this uncertainty does not necessarily mimic the uncertainty that exists among humans' perceptions about which inferences can and cannot be made.

It is worth noting that the softmax distributions tend to reflect the model's confidence on the dataset as a whole, rather than uncertainty on individual examples. For example, in the RTE2 dataset, the model nearly always splits probability mass over multiple labels, whereas in SNLI, the model typically concentrates probability mass onto a single label. This is not surprising behavior, but serves to corroborate the claim that modeling probabilistic entailment via softmax layers does

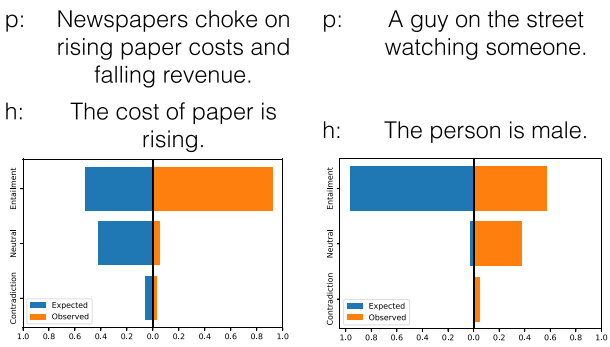

Figure 7: Examples of p/h pairs on which the model's predictions about the distribution (blue) misrepresent the nature of the uncertainty observed among human judgments (orange). In the first example (from RTE2) the model assumes ambiguity when humans consider the inference to be unambiguous (Cross-Ent = 0.36; PMF = 2.2e-6). In the second example (from SNLI) the model is certain when humans are actually in disagreement (Cross-Ent = 0.43; PMF = 5.9e-18)

not correspond to modeling annotator uncertainty about inference judgments on specific items.

6 Discussion

The results in Sections 4 and 5 suggest that 1) human NLI judgments are not adequately captured by a single aggregate score and 2) NLI systems trained to predict an aggregate score do not learn human-like models of uncertainty ''for free''. These takeaways are significant for work in computational semantics and language technology in general primarily because NLI has, historically (Cooper et al., 1996; Dagan et al., 2006) as well as presently (White et al., 2017), been proposed as a means for evaluating a model's ''intrinsic'' understanding of language: As originally framed by Dagan et al. (2006), NLI was proposed as an intermediate task for evaluating whether a model will be useful in applications, and currently, NLI is increasingly used as a means for ''probing'' neural models to assess their knowledge of arbitrary linguistic phenomena (Dasgupta et al., 2018; Ettinger et al., 2018; Poliak et al., 2018b; White et al., 2017; Poliak et al., 2018a; McCoy et al., 2019). In other words, NLI has largely become an evaluation *lingua franca* through which we diagnose what a semantic representation knows. With the increased interest in ''general-purpose'', ''task

independent'' semantic representations,[13,14] it is particularly important that intrinsic evaluations are reliable, if comparison of such representations are to be meaningful.

As discussed, the preference among many in NLP (the authors included) is to avoid tasks which take a prescriptivist approach to language and meaning. Instead, we attempt to design tasks which capture humans' linguistic behavior in as natural a setting as possible (acknowledging that truly natural annotation is difficult) with the hope that models trained to perform such tasks will be the best match for the ''real world'' settings in which we hope to deploy them. That is, we generally prefer to punt on precise definitions, and instead train our models to ''do what humans do''. In this paper, we have shown that defining ''what humans do'' is not straightforward, as humans do not necessarily handle ambiguity or communicate uncertainty in the same way as one another. Thus, as was the case for pipelined systems (Zadrozny and Elkan, 2002; Finkel et al., 2006; Bunescu, 2008) and related discussions of model calibration (Kuleshov and Liang, 2015), we argue that the best approach is to propagate uncertainty downstream, so that end tasks can decide if and how to handle inferences on which humans are likely to disagree. From the point of view of current neural NLI models—and the sentence encoders on top of which they are built—this means that a representation should be evaluated in terms of its ability to predict the full distribution of human inferences (e.g., by reporting cross-entropy against a distribution of human ratings), rather than to predict a single aggregate score (e.g., by reporting accuracy against a discrete majority label or correlation with a mean score).

We have shown that models that are trained to predict an aggregate score do not, by default, model the same type of uncertainty as that which is captured by distributions over many human raters' judgments. Thus, several challenges would need to be overcome to switch to the proposed NLI evaluation. First, NLI evaluation sets would need to be annotated by sufficiently many raters such that we can have an accurate estimate of the distribution against which to evaluate. Although the data collected for the purposes of this paper could serve as a start towards this end, a larger effort to augment or replace existing evaluation sets with full distributions of judgments would be necessary in order to yield a meaningful redefinition of the NLI task. Second, changes would be required to enable models to learn to predict these distributions. One approach could be to annotate training data, not just evaluation data, with full distributions, and optimize for the objective directly. This would clearly incur additional costs, but could be overcome with more creative crowdsourcing techniques (Dumitrache et al., 2013; Poesio et al., 2019). However, requiring direct supervision of full distributions is arguably an unsatisfying solution: Rarely if ever do humans witness multiple people responding to identical stimuli. Rather, more plausibly, we form generalizations about the linguistic phenomena that give rise to uncertainty on the basis of a large number of singly labeled examples. Thus, ideally, progress can be made by developing new architectures and/or training objectives that enable models to learn a notion of uncertainty that is consistent with the full range of possible human inferences, despite observing labels from only one or a few people on any given p/h pair. Overcoming these challenges, and moving towards models which can both understand sources of linguistic uncertainty and anticipate the range of ways that people might resolve it would be exciting both for NLI and for representation learning in general.

7 Related Work

Defining Entailment and NLI. As outlined in Section 2, there has been substantive discussion about the definition of the NLI task. This debate can largely be reduced to a debate about sentence meaning versus speaker meaning. The former aligns more closely with the goals of formal semantics and seeks a definition of the NLI task that precisely circumscribes the ways in which vague notions of ''world knowledge'' and ''common sense'' can factor into inference (Zaenen et al., 2005). The latter takes the perspective that the NLI task should maintain an informal definition in which $p \rightarrow h$ as long as h is something that a human would be ''happy to infer'' from p, where the humans making the inferences are assumed to be ''awake, careful, moderately intelligent and informed ... but not ... semanticists or similar academics'' (Manning, 2006).

[13] https://www.clsp.jhu.edu/workshops/18-workshop/general-purpose-sentence-representation-learning/

[14] https://repeval2019.github.io

Garoufi (2007) provides an overview of attempts that have been made to circumscribe the annotation process by providing finer-grained annotation options, in order to bring it more in line with the sentence-meaning task definition. Westera and Boleda (2019), in the context of advocating for distributional models of semantics in general, makes a case in favor of the speaker-meaning approach, arguing that issues like entailment, reference, and truth conditions should not fall within the purview of sentence meaning at all, despite being quintessential topics of formal semantic study. Chatzikyriakidis et al. (2017) overview NLI datasets, observing that datasets tend to be designed with one of these perspectives in mind, and thus all datasets ''fail to capture the wealth of inferential mechanisms present in NLI and seem to be driven by the dominant discourse in the field at the time of their creation.''

An orthogonal line of discussion about the definition of entailment focuses on the question of whether truth-conditional semantics should be strictly binary (propositions are either true or false) or rather treated as continuous/probabilistic values. Currently, at least within computationally minded work on textual inference, the prevailing opinion is in favor of the latter (i.e., allowing semantic judgments to be probabilistic) with few (if any) advocating that we should build systems that only support discrete true/false decisions. Still, significant theoretical and algorithmic work has gone into making probabilistic logics work in practice. Such work includes (controversial) formalisms such as fuzzy set theory (Zadeh, 1994, 1996), as well as more generally accepted formalisms which assume access to boolean groundings, such as probabilistic soft logic (Friedman et al., 1999; Kimmig et al., 2012; Beltagy et al., 2014) and Markov logic networks (Richardson and Domingos, 2006). Also related is work on collecting and analyzing graded entailment judgments (de Marneffe et al., 2012). We note that the question of strict vs. graded entailment judgments pertains to modeling of uncertainty *within* an individual rater's judgments. This is independent of the question of if/how to model disagreements *between* raters, which is the our focus in this work.

Embracing Rater Disagreement. Significant past work has looked an annotator disagreement in linguistic annotations, and has advocated that this disagreement should be taken as signal rather than noise (Aroyo et al., 2018; Palomaki et al., 2018). Plank et al. (2014) showed that incorporating rater uncertainty into the loss function for a POS tagger improves downstream performance. Similar approaches have been applied in parsing (Martínez Alonso et al., 2015) and supersense tagging (Martínez Alonso et al., 2016). Specifically relevant to this work is past discussion of disagreement on semantic annotation tasks, including anaphora resolution (Poesio and Artstein, 2005), coreference (Versley, 2008; Recasens et al., 2011), word sense disambiguation (Erk and McCarthy, 2009; Passonneau et al., 2012; Jurgens, 2013), veridicality (Geis and Zwicky, 1971; Karttunen et al., 2014; de Marneffe et al., 2012), semantic frames (Dumitrache et al., 2019), and grounding (Reidsma and op den Akker, 2008).

Most of this work focuses on the uncertainty of individual raters, oftentimes concluding that such uncertainty can be addressed by shifting to a graded rather than discrete labeling schema and/or that uncertainty can be leveraged as a means for detecting inherently ambiguous items. In contrast, we do not look at measures of uncertainty/ambiguity from the point of view of an individual (though this is a very interesting question); rather, we focus on disagreements that exist between raters. We agree strongly that semantic judgments should be treated as graded, and that ambiguous items should be acknowledged as such. Still, this is independent of the issue of inter-rater disagreement: Two raters can disagree when making graded judgments as much as they can when making discrete judgments, and they can disagree when they are both uncertain as much as they can when they are both certain. Thus, the central question of this work is whether aggregation (via average or majority vote) is a faithful representation of the underlying distribution of judgments across annotators. Arguably, such aggregation is a faithful (albeit lossy) representation of high-variance unimodal distributions, but not of multi-modal ones.

In this regard, particularly relevant to our work is de Marneffe et al. (2012) and de Marneffe et al. (2018), who observed similarly persistent disagreement in graded judgments of veridicality, and made a case for attempting to model the full distribution as opposed to a single aggregate score. Smith et al. (2013) present related theoretical work, which proposes specific mechanisms by

which humans might handle lexical uncertainty in the context of inference. Their model assumes pragmatic speakers and listeners who reason simultaneously about one another's goals and about the lexicon itself, and could be used to explain differing inferences in cases where raters share different beliefs about the speaker (author) of p and/or about the lexicon. Schaekermann et al. (2016) develop a proof-of-concept annotation interface specifically intended to recognize whether or not inter-rater disagreement is ''resolvable'' via more annotation, or rather is likely to persist, although they don't discuss natural language semantics directly. Finally, Tanenhaus et al. (1985) discuss the role of formal semantics and generative grammar in inference, and specifically differentiates between work which treats grammar as a causal process of how inferences occur versus work which treats grammar as a descriptive framework of the structure of language. Such discussion is relevant going forward, as engineers of NLI systems must determine both how to define the evaluation task, as well as the role that concepts from formal semantics should play within such systems.

8 Conclusion

We provide an in-depth study of disagreements in human judgments on the NLI task. We show that many disagreements persist even after increasing the number of annotators and the amount of context provided, and that models which represent these annotations as multimodal distributions generalize better to held-out data than those which do not. We evaluate whether a state-of-the-art NLI model (BERT) captures these disagreements by virtue of producing softmax distributions over labels and show that it does not. We argue that, if NLI is to serve as an adequate intrinsic evaluation of semantic representations, then models should be evaluated in terms of their ability to predict the full expected distribution over all human raters, rather than a single aggregate score.

Acknowledgments

Thank you to the Action Editor Chris Potts and the anonymous reviewers for their input on earlier drafts of this paper. This work evolved substantially as a result of their suggestions and feedback. Thank you to Dipanjan Das, Michael Collins, Sam Bowman, Ankur Parikh, Emily Pitler, Yuan Zhang, and the rest of the Google Language team for many useful discussions.

References

Lora Aroyo, Anca Dumitrache, Praveen Paritosh, Alex Quinn, and Chris Welty, editors. 2018. *Proc. Subjectivity, Ambiguity and Disagreement in Crowdsourcing (SAD)*, volume 1 of *1*. HCOMP, Zurich, Switzerland.

Islam Beltagy, Katrin Erk, and Raymond Mooney. 2014. Probabilistic soft logic for semantic textual similarity. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1210–1219, Baltimore, MD. Association for Computational Linguistics.

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.

Razvan Bunescu. 2008. Learning with probabilistic features for improved pipeline models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 670–679, Honolulu, HI. Association for Computational Linguistics.

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with Amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12, Los Angeles, CA. Association for Computational Linguistics.

Stergios Chatzikyriakidis, Robin Cooper, Simon Dobnik, and Staffan Larsson. 2017. An overview of natural language inference data collection: The way forward? In *Proceedings of the Computing Natural Language Inference Workshop*.

Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the Web for fine-grained semantic verb relations. In *Proceedings of*

the 2004 Conference on Empirical Methods in Natural Language Processing, pages 33–40, Barcelona, Spain. Association for Computational Linguistics.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, and Steve Pullman. 1996, Using the framework. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006, The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, pages 177–190. Springer

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J. Gershman, and Noah D. Goodman. 2018. Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN. Association for Computational Linguistics.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2019. A crowdsourced frame disambiguation corpus with ambiguity. 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 2164–2170, Minneapolis, MN. Association for Computational Linguistics.

Anca Dumitrache, Lora Aroyo, Christopher A. Welty, Robert-Jan Sips, and Anthony Levas. 2013. Dr. Detective: Combining gamification techniques and crowdsourcing to create a gold standard for the medical domain. In *CrowdSem Workshop at the International Semantic Web Conference*.

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.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801, Santa Fe, NM. Association for Computational Linguistics.

Christiane Fellbaum. 1998. WordNet, Wiley Online Library.

Jenny Rose Finkel, Christopher D. Manning, and Andrew Y. Ng. 2006. Solving the problem of cascading errors: Approximate Bayesian inference for linguistic annotation pipelines. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 618–626, Sydney, Australia. Association for Computational Linguistics.

Nir Friedman, Lise Getoor, Daphne Koller, and Avi Pfeffer. 1999. Learning probabilistic relational models. In *IJCAI*, 99: 1300–1309.

Konstantina Garoufi. 2007. Towards a Better Understanding of Applied Textual Entailment. Ph.D. thesis, Citeseer.

Michael L. Geis and Arnold M. Zwicky. 1971. On invited inferences. *Linguistic inquiry*, 2(4):561–566.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics*.

Rafal Józefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *ArXiv*, abs/1602.02410.

David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 556–562, Atlanta, GA. Association for Computational Linguistics.

Lauri Karttunen, Stanley Peters, Annie Zaenen, and Cleo Condoravdi. 2014. The chameleon-like nature of evaluative adjectives. *Empirical Issues in Syntax and Semantics*, 10:233–250.

Angelika Kimmig, Stephen Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2012. A short introduction to probabilistic soft logic. In *Proceedings of the NIPS Workshop on Probabilistic Programming: Foundations and Applications*, pages 1–4.

Volodymyr Kuleshov and Percy S. Liang. 2015. Calibrated structured prediction. In *Advances in Neural Information Processing Systems*, pages 3474–3482.

Christopher D. Manning. 2006. Local textual inference: Its hard to circumscribe, but you know it when you see it–and NLP needs it. https://nlp.stanford.edu/manning/papers/TextualInference.pdf

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2018. Factivity in doubt: Clause-embedding predicates in naturally occurring discourse. *Sinn und Bedeutung 23 (Poster)*.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

Héctor Martínez Alonso, Anders Johannsen, and Barbara Plank. 2016. Supersense tagging with inter-annotator disagreement. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 43–48, Berlin, Germany. Association for Computational Linguistics.

Héctor Martínez Alonso, Barbara Plank, Arne Skjærholt, and Anders Søgaard. 2015. Learning to parse with IAA-weighted loss. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1357–1361, Denver, CO. Association for Computational Linguistics.

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.

Richard Montague. 1970. Universal grammar. *Theoria*, 36(3):373–398.

Jennimaria Palomaki, Olivia Rhinehart, and Michael Tseng. 2018. A case for a range of acceptable annotations. In *Proceedings of Workshop on Subjectivity, Ambiguity, and Disagreement (SAD)*. HCOMP.

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.

Ellie Pavlick and Chris Callison-Burch. 2016a. Most ''babies'' are ''little'' and most ''problems'' are ''huge'': Compositional entailment in adjective-nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2164–2173, Berlin, Germany. Association for Computational Linguistics.

Ellie Pavlick and Chris Callison-Burch. 2016b. So-called non-subsective adjectives. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 114–119, Berlin, Germany. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. 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.

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, MI. Association for Computational Linguistics.

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, MN. Association for Computational Linguistics.

Adam Poliak, Yonatan Belinkov, James Glass, and Benjamin Van Durme. 2018a. On the evaluation of semantic phenomena in neural machine translation using natural language inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 513–523, New Orleans, LA. Association for Computational Linguistics.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018b. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium. Association for Computational Linguistics.

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.

Dennis Reidsma and Rieks op den Akker. 2008. Exploiting 'subjective' annotations. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 8–16, Manchester, UK. Coling 2008 Organizing Committee.

Matthew Richardson and Pedro Domingos. 2006. Markov logic networks. *Machine Learning*, 62(1–2):107–136.

Mike Schaekermann, Edith Law, Alex C. Williams, and William Callaghan. 2016. Resolvable vs. irresolvable ambiguity: A new hybrid framework for dealing with uncertain ground truth. In *1st Workshop on Human-Centered Machine Learning at SIGCHI*.

Mandy Simons, Judith Tonhauser, David Beaver, and Craige Roberts. 2010. What projects and why. *Semantics and Linguistic Theory*, 20: 309–327.

Nathaniel J. Smith, Noah Goodman, and Michael Frank. 2013. Learning and using language via recursive pragmatic reasoning about other agents. In *Advances in Neural Information Processing Systems*, pages 3039–3047.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, HI. Association for Computational Linguistics.

M. Tanenhaus, G. Carlson, and M. S. Seidenberg. 1985. Do listeners compute linguistic representations? D. Dowty, L. Karttunen, and A. Zwicky, editors, *Natural Language Parsing*. Cambridge University Press.

Judith Tonhauser, David I. Beaver, and Judith Degen. 2018. How Projective is Projective Content? Gradience in Projectivity and At-issueness. *Journal of Semantics*, 35(3):495–542.

Y. Versley. 2008. Vagueness and referential ambiguity in a large-scale annotated corpus. *Research on Language and Computation*, 6:333–353.

Matthijs Westera and Gemma Boleda. 2019. Don't blame distributional semantics if it can't do entailment. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 120–133, Gothenburg, Sweden. Association for Computational Linguistics.

Aaron S. White, Valentine Hacquard, and Jeffrey Lidz. 2018. Semantic information and the syntax of propositional attitude verbs. *Cognitive Science*, 42(2):416–456.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005,

Taipei, Taiwan. Asian Federation of Natural Language Processing.

Aaron Steven White and Kyle Rawlins. 2017. The role of veridicality and factivity in clause selection. *48th Annual Meeting of the North East Linguistic Society*, Reykjavík. http://iceland2017.nelsconference.org/wp-content/uploads/2017/08/White-Rawlins.pdf.

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, LA. Association for Computational Linguistics.

Lotfi A. Zadeh. 1994. Fuzzy logic, neural networks, and soft computing. *Communications of the ACM*, 37(3):77–84.

Lotfi A. Zadeh. 1996. Fuzzy logic = computing with words. *IEEE Transactions on Fuzzy Systems*, 4(2):103–111.

Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699. ACM.

Annie Zaenen, Lauri Karttunen, and Richard Crouch. 2005. Local textual inference: Can it be defined or circumscribed? In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 31–36, Ann Arbor, MI. Association for Computational Linguistics.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 1996. Ordinal Common-sense Inference. *Transactions of the Association for Computational Linguistics*, 5:379–395. https://www.aclweb.org/anthology/Q17-1027. doi:10.1162/tac_a_00068.