Uncertain Natural Language Inference

Tongfei Chen\textsuperscript{1\ast} Zhengping Jiang\textsuperscript{2\ast\dagger} Adam Poliak\textsuperscript{1} 
Keisuke Sakaguchi\textsuperscript{3\dagger} Benjamin Van Durme\textsuperscript{1}

\textsuperscript{1} Johns Hopkins University \hspace{1cm} \textsuperscript{2} Columbia University \hspace{1cm} \textsuperscript{3} Allen Institute for AI

{tongfei, azpoliak, vandurme}@jhu.edu, zj2265@columbia.edu, keisukes@allenai.org

Abstract

We introduce Uncertain Natural Language Inference (UNLI), a refinement of Natural Language Inference (NLI) that shifts away from categorical labels, targeting instead the direct prediction of subjective probability assessments. We demonstrate the feasibility of collecting annotations for UNLI by relabeling a portion of the SNLI dataset under a probabilistic scale, where items even with the same categorical label differ in how likely people judge them to be true given a premise. We describe a direct scalar regression modeling approach, and find that existing categorically labeled NLI data can be used in pre-training. Our best models approach human performance, demonstrating models may be capable of more subtle inferences than the categorical bin assignment employed in current NLI tasks.

1 Introduction

Variants of entailment tasks have been used for decades in benchmarking systems for natural language understanding. Recognizing Textual Entailment (RTE) or Natural Language Inference (NLI) is traditionally a categorical classification problem: predict which of a set of discrete labels apply to an inference pair, consisting of a premise ($p$) and hypothesis ($h$). The FraCaS consortium offered the task as an evaluation mechanism, along with a small challenge set (Cooper et al., 1996), which was followed by the RTE challenges (Dagan et al., 2005). Despite differences between these and recent NLI datasets (Marelli et al., 2014; Lai et al., 2017; Williams et al., 2018; Khot et al., 2018, i.a.), NLI has remained a categorical prediction problem.

However, entailment inference is uncertain and has a probabilistic nature (Glickman et al., 2005). Maintaining NLI as a categorical classification problem is not ideal since coarse categorical labels mask the uncertain and probabilistic nature of entailment inference. NLI pairs may share a coarse label, but the probabilities that the hypotheses are entailed by their corresponding premises may vary greatly (see Table 1). Hence, not all contradictions are equally contradictory and not all entailments are equally entailed.

We propose Uncertain Natural Language Inference (UNLI), a refinement of NLI that captures more subtle distinctions in meaning by shifting away from categorical labels to the direct prediction of human subjective probability assessments. We illustrate that human-elicited probability assessments contain subtle distinctions on the likelihood of a hypothesis conditioned on a premise, and UNLI captures these distinctions far beyond categorical labels in popular NLI datasets.

We demonstrate how to elicit UNLI annotations. Using recent large-scale language model pre-training, we provide experimental results illustrating that systems can often predict UNLI judgments, but with clear gaps in understanding. We conclude that scalar annotation protocols should be adopted in future NLI-style dataset creation, which should enable new work in modeling a richer space of interesting inferences.

<table>
<thead>
<tr>
<th>Premise $\rightarrow$ Hypothesis</th>
<th>NLI</th>
<th>UNLI</th>
</tr>
</thead>
<tbody>
<tr>
<td>A man in a white shirt taking a picture</td>
<td>ENT</td>
<td>100%</td>
</tr>
<tr>
<td>A man takes a picture</td>
<td>ENT</td>
<td>100%</td>
</tr>
<tr>
<td>A boy hits a ball, with a bat</td>
<td>ENT</td>
<td>78%</td>
</tr>
<tr>
<td>The kid is playing in a baseball game</td>
<td>ENT</td>
<td>78%</td>
</tr>
<tr>
<td>A wrestler in red cries, one in blue celebrates</td>
<td>CON</td>
<td>50%</td>
</tr>
<tr>
<td>The wrestler in blue is undefeated</td>
<td>CON</td>
<td>50%</td>
</tr>
<tr>
<td>Man laying on a platform outside on rocks</td>
<td>CON</td>
<td>0%</td>
</tr>
<tr>
<td>Man takes a nap on his couch</td>
<td>CON</td>
<td>0%</td>
</tr>
</tbody>
</table>

Table 1: Probability assessments on NLI pairs. The NLI and UNLI columns respectively indicate the categorical label (from SNLI) and the subjective probability for the corresponding pair.
We elicit subjective probabilities from crowdsource workers (MTurk) for premise-hypothesis pairs from existing NLI data. Annotators are asked to estimate how likely the situation described in the hypothesis sentence would be true given the premise. Following the Efficient Annotation of Scalar Labels framework (EASL; Sakaguchi and Durme, 2018), we present annotators 5 sentence-pairs, each with a slider bar enabling direct assessment for each pair and ask annotators to calibrate their score for a sentence-pair based on the scores they provided to the other four pairs.1

In contrast to the uniform scale employed in the original EASL protocol, we modify the interface to allow finer-grained values near 0.0 and 1.0, following psychological findings that humans are especially sensitive to values near the ends of the probability spectrum (Tversky and Kahneman, 1981). This interface decision is a key distinction of this work contrasting prior efforts that averaged Likert-scale (ordinal) annotations. This allows us to capture the difference between NLI pairs that are both appropriately contradicted or entailed under NLI, but that have a perceived difference of less than 1% probability.

In order to capture the sensitivity near these ends, we adopt a more fine-grained slider bar with 10,000 steps with a logistic transformation. Specifically, for raw score \( x \in [0, 10000] \), we apply a scaled logistic function \( f(x) = \sigma(\beta(x - 5000)) \) to re-scale the final result range to \([0, 1]\). We ran pilots to tune \( \beta \), and determine that people tend to choose much lower probability for some events even though they are just slightly less likely (e.g., just below 50%).

Therefore, we use different \( \beta \)'s depending on the range of \([0, 0.5]\) or \((0.5, 1]\). Each sentence pair is annotated with 2- or 3-way redundancy. The individual responses are averaged to create a gold standard label for a premise-hypothesis pair.

### Data
We annotate, i.e., elicit a probability \( y \in [0, 1] \), for a subset of SNLI (Bowman et al., 2015) examples and refer to this data as \( \mu\text{-SNLI} \). \( \mu\text{-SNLI} \)'s training set contains 7,931 distinct premises paired with at least 5 distinct neutral (NEU) hypotheses. For each premise, we sample 5 neutral hypotheses, resulting in 39,655 of these NEU pairs annotated.

An additional 15,862 contradicted (CON) and entailed (ENT) pairs are annotated for our training set, resulting in 55,517 training examples. For our dev and test sets, we respectively annotated 3,040 examples from SNLI’s dev and test sets. In total, we annotated 61,597 examples, about 12% of all examples in SNLI. Figure 1 plots the resultant median and quartile for each of the 3 categories under our scalar probability scheme. Light / dark shade covers 96% / 50% of each category, and the bar denotes the median. Note that x-axis is logistic to allow fine-grained distinctions near 0.0 and 1.0.

Therefore, we use different \( \beta \)'s depending on the range of \([0, 0.5]\) or \((0.5, 1]\). Each sentence pair is annotated with 2- or 3-way redundancy. The individual responses are averaged to create a gold standard label for a premise-hypothesis pair.

![Figure 1: Dev set statistics, illustrating median and quartile for each of the 3 categories under our scalar probability scheme. Light / dark shade covers 96% / 50% of each category, and the bar denotes the median. Note that x-axis is logistic to allow fine-grained distinctions near 0.0 and 1.0.](image)

### Table 2: A premise in SNLI with its 5 hypotheses (labeled as neutral in SNLI) annotated in \( \mu\text{-SNLI} \).

<table>
<thead>
<tr>
<th>Premise</th>
<th>Hypothesis</th>
<th>SNLI</th>
<th>( \mu\text{-SNLI} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>~→ A man performs a song.</td>
<td>NEU</td>
<td>95%</td>
<td></td>
</tr>
<tr>
<td>~→ A man is performing on stage.</td>
<td>NEU</td>
<td>84%</td>
<td></td>
</tr>
<tr>
<td>A man is singing into a microphone. ~→ A male performer is singing a special and meaningful song.</td>
<td>NEU</td>
<td>15%</td>
<td></td>
</tr>
<tr>
<td>~→ A man performing in a bar.</td>
<td>NEU</td>
<td>14%</td>
<td></td>
</tr>
<tr>
<td>~→ A man is singing the national anthem at a crowded stadium.</td>
<td>NEU</td>
<td>0.6%</td>
<td></td>
</tr>
</tbody>
</table>

1 Example pairs were provided in the instructions along with suggested probability values. See Appendix A for details of the annotation interface and qualifications.

2 This is called the certainty effect: more sensitivity to the difference between, e.g., 0% and 1% than 50% and 51%.

3 This phenomenon accords with the weighting function in Prospect Theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992), where people tend to downweight probabilities with around 0.4 or above.

4 We use SNLI due to its popularity and its feature that each premise is paired with multiple hypotheses.

5 Data is available at [http://nlp.jhu.edu/unli](http://nlp.jhu.edu/unli).
We train a regression UNLI model to predict the probability that a premise entails a hypothesis. We modify the sentence pair classifier\(^6\) in BERT to exploit recent advancements in large-scale language model pre-training. Following Devlin et al. (2019), we concatenate the premise and the hypothesis, with a special sentinel token (CLS) inserted at the beginning and a separator (SEP) inserted after each sentence, tokenized using WordPiece. After encoding the concatenated token sequence with BERT, we take the encoding of the first sentinel token.

\[ f(p, h) = \text{BERT}(\text{CLS}; p; \text{SEP}; h; \text{SEP})[0]. \]

We pass the resulting feature vector \( f(p, h) \) through a sigmoid-activated linear layer to obtain a probability, instead of a softmax used in categorical NLI. We directly model UNLI as a regression problem, trained using a binary cross-entropy loss. Pearson \( r \) measures the linear correlation between the gold probability assessments and model’s output; Spearman \( \rho \) measures the ability of the model ranking the premise-hypothesis pairs with respect to their subjective probability; MSE measures whether the model can recover the subjective probability value from premise-hypothesis pairs. A high \( r \) and \( \rho \), but a low MSE is desired.

### 4 Results & Analysis

Table 4 reports results on \( u \)-SNLI dev and test sets. Just training on 55,517 \( u \)-SNLI examples yields a 62.71\% Pearson \( r \) on test. The hypothesis-only baseline achieved a correlation around 40\%. This result corroborates the findings that a hidden bias exists in the SNLI dataset’s hypotheses, and shows this bias may also exist in \( u \)-SNLI.\(^9\)

<table>
<thead>
<tr>
<th>Hyp-only</th>
<th>Full-model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev</td>
<td>Test</td>
</tr>
<tr>
<td>( r )</td>
<td>0.3759</td>
</tr>
<tr>
<td>( \rho )</td>
<td>0.3853</td>
</tr>
<tr>
<td>MSE</td>
<td>0.1086</td>
</tr>
</tbody>
</table>

Table 4: Metrics for training on \( u \)-SNLI.

**Human Performance** We elicit additional annotations on \( u \)-SNLI dev set to establish a randomly sampled human performance. We use the same annotators as before but ensure each annotator has not previously seen the pair they are annotating. We average the scores from three-way redundant elicitation,\(^10\) yielding \( r = 0.6978, \rho = 0.7273 \), and MSE = 0.0759: our regression model trained on \( u \)-SNLI is therefore approaching human performance. While encouraging, the model fails drastically for some examples.

---

\(^6\) The neural architecture for MultiNLI (Williams et al., 2018) in Devlin et al. (2019).

\(^7\) No significant difference is observed with an \( L_2 \) loss.

\(^8\) See Appendix D for additional training details.

\(^9\) This is unsurprising because \( u \)-SNLI examples are sampled from SNLI.

\(^10\) This setting approximates the performance of a randomly sampled human on \( u \)-SNLI, and is therefore a reasonable lower bound on the performance one could achieve with a dedicated, trained single human annotator.
Qualitative Error Analysis  Table 3 illustrates examples with large gaps between the gold probability assessment and the BERT-based model output. The model seems to have learned lexicon-level inference (e.g., race cars $\Rightarrow$ going fast, but ignored crucial information (sits in the pits), and fails to learn certain commonsense patterns (e.g. riding amusement park ride $\Rightarrow$ screaming; man and woman drinking at a bar $\Rightarrow$ on a date). These examples illustrate the model’s insufficient commonsense reasoning and plausibility estimation.

Pre-training with SNLI  Can we leverage the remaining roughly 500,000 SNLI training pairs that only have categorical labels? One method would be to train a categorical NLI model on SNLI and when fine-tuning on $u$-SNLI, replace the last layer of the network from a categorical prediction with a sigmoid function.\(^\text{11}\) However, a typical categorical loss function would not take into account the ordering between the different categorical labels.\(^\text{12}\) Instead, we derive a surrogate function $s : T \rightarrow [0, 1]$ that maps SNLI categorical labels $t \in \{ \text{ENT, NEU, CON} \}$ to the average score of all $u$-SNLI training annotations labeled with $t$ in SNLI.\(^\text{13}\)

<table>
<thead>
<tr>
<th>SNLI</th>
<th>$u$-SNLI</th>
<th>Dev</th>
<th>Test</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r$</td>
<td>0.5198</td>
<td>0.4958</td>
<td>0.6762</td>
<td>0.6589</td>
<td></td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.5238</td>
<td>0.5231</td>
<td>0.6806</td>
<td>0.6708</td>
<td></td>
</tr>
<tr>
<td>MSE</td>
<td>0.1086</td>
<td>0.0928</td>
<td>0.0694</td>
<td>0.0733</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Metrics for training only on mapped SNLI or fine-tuning on $u$-SNLI.

We use this mapping to pre-train a regression model on the SNLI training examples not included in $u$-SNLI. We also fine-tune the model on $u$-SNLI’s training set. Table 5 reports the results evaluated on $u$-SNLI’s dev and test sets. The model trained on the roughly 500K mapped SNLI examples, performs much worse than when trained on just about 55K $u$-SNLI examples. When we pre-train the model on the mapped SNLI and fine-tune on $u$-SNLI, results noticeably improve. This improvement is akin to the Phang et al. (2018)’s finding that many NLI datasets cover informative signal for different tasks, explaining why pre-training on NLI can be advantageous. Here, an impoverished version of UNLI is helpful.

Model behavior  Figure 2 depicts the model behavior when training just on SNLI or fine-tuning with $u$-SNLI. When using the original SNLI data, under the surrogate regression setting, the model’s prediction concentrates on the 3 surrogate scalar values of the 3 SNLI classes. After fine-tuning on $u$-SNLI, the model learns smoother predictions for premise-hypothesis pairs, supported by the superior Pearson correlation score. The darker boxes in bottom-right corner of the heatmaps (Figure 2) indicate high accuracy on samples with $\approx 1.0$ gold $u$-SNLI labels and $\approx 1.0$ model predictions, signifying that our UNLI models are very good at recognizing entailments.

Figure 2: Heatmap on $u$-SNLI dev predictions when trained only on SNLI (left) or fine-tuned on $u$-SNLI (right). Prediction frequencies are normalized along each gold label row.

5 Related Work

The probabilistic nature and the uncertainty of NLI has been considered from a variety of perspectives. Glickman et al. (2005) modified the task to explicitly include the probabilistic aspect of NLI, stating that “$p$ probabilistically entails $h \ldots$ if $p$ increases the likelihood of $h$ being true,” while Lai and Hockenmaier (2017) noted how predicting the conditional probability of one phrase given another would be helpful in predicting textual entailment. Other prior work has elicited ordinal annotations (e.g. Likert scale) reflecting likelihood judgments (Pavlick and Callison-Burch, 2016; Zhang et al., 2017), but then collapsed the annotations into coarse categorical labels for modeling. Vulić et al. (2017) proposed graded lexical entailment, which is similar to our idea but applied to lexical-level inference, asking “to what degree $x$ is a type of $y$.” Additionally, Lalor et al. (2016, 2018) tried capturing the uncertainty of each inference pair by item response theory (IRT), showing fine-grained...
differences in discriminative power in each label. Pavlick and Kwiatkowski (2019) recently argued that models should “explicitly capture the full distribution of plausible human judgments” as plausible human judgments cause inherent disagreements. Our concern is different as we are interested in the uncertain and probabilistic nature of NLI. We are the first to propose a method for direct elicitation of subjective probability judgments on NLI pairs and direct prediction of these scalars, as opposed to reducing to categorical classification.

Recent work have also modeled the uncertainty of other semantic phenomena as direct scalar regression (and collected scalar versions of data for them) instead of categorical classification, e.g. factuality (Lee et al., 2015; Stanovsky et al., 2017; Rudinger et al., 2018), and semantic proto-roles (Teichert et al., 2017).

Plausibility tasks such as COPA (Roemmele et al., 2011) and ROCStories (Mostafazadeh et al., 2016) ask models to choose the most probable examples given a context, capturing relative uncertainty between examples, but do not force a model to predict the probability of $h$ given $p$. Li et al. (2019) viewed the plausibility task of COPA as a learning to rank problem, where the model is trained to assign the highest scalar score to the most plausible alternative given context. Our work can be viewed as a variant to this, with the score being an explicit human probability judgment instead.

Linguists such as van Eijck and Lappin (2014), Goodman and Lassiter (2015), Cooper et al. (2015) and Bernardy et al. (2018) have described models for natural language semantics that introduce probabilities into the compositional, model-theoretic tradition begun by those such as Davidson (1967) and Montague (1973). Where they propose probabilistic models for interpreting language, we are concerned with illustrating the feasibility of eliciting probabilistic judgments on examples through crowdsourcing, and contrasting with prior efforts restricted to limited categorical label sets.

6 Conclusion

We proposed Uncertain Natural Language Inference (UNLI), a new task of directly predicting human likelihood judgments on NLI premise-hypothesis pairs. In short, we have shown that not all NLI contradictions are created equal, nor neutrals, nor entailments. We demonstrated that (1) eliciting supporting data is feasible, and (2) annotations in the data can be used for improving a scalar regression model beyond the information contained in existing categorical labels, using recent contextualized word embeddings, e.g. BERT.

Humans are able to make finer distinctions between meanings than is being captured by current annotation approaches; we advocate the community strives for systems that can do the same, and therefore shift away from categorical NLI labels and move to something more fine-grained such as our UNLI protocol.

Acknowledgments

We thank anonymous reviewers from current and past versions of the article for their insightful comments and suggestions. This research benefited from support by DARPA AIDA and DARPA LORELEI. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.

References


Paulo Galeazzi, Nina Gierasimeczuk, Alexandru Marcoci, and Sonja Smets, editors, The Logic and Interactive Rationality Yearbook 2012, volume II.


Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. The SICK (Sentences Involving Compositional Knowledge) dataset for relatedness and entailment.


A Annotation

Here we include information about the qualifications used to vet annotators. We also include screenshots of the interface used to collect annotations.

A.1 Qualification Test

Annotators were given a qualification test to ensure non-expert workers were able to give reasonable subjective probability estimates. We first extracted seven statements from Book of Odds (Shapiro et al., 2014), and manually split the statement into a bleached premise and hypothesis. We then wrote three easy premise-hypothesis pairs with definite probabilities like ($p = "A girl tossed a coin.\), $h = "The coin comes up a head.\)”, probability: 0.5). We qualify users that meet both criteria: (1) For the three easy pairs, their annotations had to fall within a small error range around the correct label $H$, computed as $\frac{1}{4} \min \{ H, 1 - H \}$. (2) Their overall annotations have a Pearson $r > 0.7$ and Spearman $\rho > 0.4$. This qualification test led to a pool of 40 trusted annotators, which were employed for the entirety of our dataset creation.

A.2 Annotation Interface

We include screenshots of the instructions and examples shown to crowdsourced workers (Figure 4) as the interface we provided (Figure 3).

B Redundant Annotations

By default, we use two crowdsourced workers to annotate each UNLI sentence-pair. If the two annotations on the raw slider bar $\{0, \ldots, 10000\}$ differ by more than 2000, we then elicit a third annotator.

C Dataset Statistics

Table 6 summarizes the statistics of $u$-SNLI.

D Additional Training Details

We use the BERT-BASE-UNCASED model, with the Adam optimizer (Kingma and Ba, 2015), an initial learning rate of $10^{-5}$, and maximum gradient norm 1.0. Our model is trained for 3 epochs, where the epoch resulting in the highest Pearson $r$ on the dev set is selected.

<table>
<thead>
<tr>
<th>Partition</th>
<th>Breakdown</th>
<th>SNLI</th>
<th>U-SNLI</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td>Distinct premises</td>
<td>151k</td>
<td>7,931</td>
</tr>
<tr>
<td></td>
<td>ENT hypotheses</td>
<td>183k</td>
<td>7,931</td>
</tr>
<tr>
<td></td>
<td>NEU hypotheses</td>
<td>183k</td>
<td>39,655</td>
</tr>
<tr>
<td></td>
<td>CON hypotheses</td>
<td>183k</td>
<td>7,931</td>
</tr>
<tr>
<td></td>
<td>Total P-H pairs</td>
<td>550k</td>
<td>55,517</td>
</tr>
<tr>
<td>dev</td>
<td>Distinct premises</td>
<td>3,319</td>
<td>2,647</td>
</tr>
<tr>
<td></td>
<td>ENT hypotheses</td>
<td>3,329</td>
<td>162</td>
</tr>
<tr>
<td></td>
<td>NEU hypotheses</td>
<td>3,235</td>
<td>2,764</td>
</tr>
<tr>
<td></td>
<td>CON hypotheses</td>
<td>3,278</td>
<td>114</td>
</tr>
<tr>
<td></td>
<td>Total P-H pairs</td>
<td>10k</td>
<td>3,040</td>
</tr>
<tr>
<td>test</td>
<td>Distinct premises</td>
<td>3,323</td>
<td>2,635</td>
</tr>
<tr>
<td></td>
<td>ENT hypotheses</td>
<td>3,368</td>
<td>156</td>
</tr>
<tr>
<td></td>
<td>NEU hypotheses</td>
<td>3,219</td>
<td>2,770</td>
</tr>
<tr>
<td></td>
<td>CON hypotheses</td>
<td>3,237</td>
<td>114</td>
</tr>
<tr>
<td></td>
<td>Total P-H pairs</td>
<td>10k</td>
<td>3,040</td>
</tr>
</tbody>
</table>

Table 6: Statistics of SNLI data re-annotated under UNLI.