WANLI - Collaboration for NLI Dataset Creation

May 16, 2022
Alisa Liu, Swabha Swayamdipta, Noah Smith, Yejin Choi
Datasets in natural language processing

Datasets are the backbone of machine learning

good training sets teach our model the task

good test sets evaluate progress

Training dataset

How can we distill human linguistic competence into datasets that models can learn from and be evaluated on?
An animal sat on the mat.

A fluffy cat sat on the mat.

A cat sat on the mat.

No one sat on the mat.

Natural language inference

Hypotheses

Premise

Replace specific words with general ones

Just negate it!

Add a plausible adjective

Models overfit to these patterns and don’t produce the right answer for the right reasons

(Geva et al., 2021; Gururangan et al., 2018)
Idea: our linguistic competence is largely subconscious

When can you replace “want to” with “wanna”?

This is the man who I want to die.

* This is the man who I wanna die.

(Lakoff, 1970)

The most subtle & “human” parts of our language understanding are largely inaccessible to us
How should this affect the way we collect data?

Humans are not good at painting a complete picture of what we know how to do with language

But we are good at evaluating what’s right and what’s wrong!

💡 We want to use humans to revise + evaluate examples... but where can we get decent examples to start with?

This is where AI comes in!
What is good at?

Humans are reliable for writing correct examples and evaluating examples

What is good at?

Large LMs are producing increasingly human-like text (Clark et al., 2021), being deployed in creative applications (Lee et al., 2022), and can replicate a pattern given just a few examples in-context (Brown et al., 2020).
Worker-AI Collaboration

Leverage the **generative strength** of LMs and **evaluative strength** of humans

LMs create new examples by replicating valuable reasoning patterns in an existing dataset

Humans revise and assign a label

We don’t expect humans to be creative, just thoughtful. We don’t expect models to be intelligent, just fluent.
Approaches to Dataset Creation
adapted from Bowman and Dahl, 2021

Crowdsourcing
(Mihaylov et al., 2018; Rajpurkar et al., 2018; Bowman et al., 2013)

Naturally-occurring examples

Expert-authored examples

Adversarial data collection

Fully generated examples

Task instructions

😍 Flexible
😍 Scalable
🙁 Annotation artifacts
Approaches to Dataset Creation
adapted from Bowman and Dahl, 2021

Crowdsourcing

Naturally-occurring examples
(Kwiatkowski et al., 2019, Narayan et al., 2018)

Expert-authored examples

Adversarial data collection

Fully generated examples

- May not exist for the desired task
- Tied to the use contexts of a specific NLP product

naturally-occurring questions!

naturally-occurring summaries!
Approaches to Dataset Creation
adapted from Bowman and Dahl, 2021

Crowdsourcing

Naturally-occurring examples

Expert-authored examples
(Levesque et al., 2012; Wang et al., 2019)

Adversarial data collection

Fully generated examples

😍 Challenging
😊 Not fitting for a broad-coverage dataset
😔 Not scalable
Approaches to Dataset Creation

adapted from Bowman and Dahl, 2021

Crowdsourcing

Naturally-occurring examples

Expert-authored examples

Adversarial data collection
(Kiela et al., 2021; Le Bras et al., 2020; Wallace et al., 2021)

Fully generated examples

มากๆ ไม่สามารถทำได้

What about this one?

I can do this already!

What about this one?

Humans better explore reasoning space

Greater annotator effort
(Bartolo et al., 2020)

May not lead to better generalization on non-adversarial test sets
(Kaushik et al., 2021)

Depends greatly on the adversaries used
(Phang et al., 2021; Zellers et al., 2019)

May result in examples beyond the scope of the task
Approaches to Dataset Creation

adapted from Bowman and Dahl, 2021

Crowdsourcing

Naturally-occurring examples

Expert-authored examples

Adversarial data collection

Fully generated examples

(Schick & Schütze, 2021; West et al., 2021; Puri et al., 2020; Lee et al., 2021)

😊 No human effort
🤔 Complexity of examples is limited to what is accessible by the model
Approaches to Dataset Creation
adapted from Bowman and Dahl, 2021

Crowdsourcing

Naturally-occurring examples

Expert-authored examples

Adversarial data collection

Fully generated examples

Worker-AI collaboration

this work!

😊 Use LM to explore more reasoning
😊 Human review ensures quality and validity
😊 Lower annotator effort
The Task: Natural Language Inference

Determine whether a piece of text entails, contradicts, or is neutral to another piece of text

Annotation artifacts in NLI datasets are well-studied (Gururangan et al., 2018, McCoy et al., 2019)

Extremely resource-available but still far from solved

Has the potential to be useful in downstream applications (Chen et al., 2021, Goyal et al., 2020)
automatically collect pockets of examples that exemplify valuable reasoning patterns

leverage GPT-3 to generate new examples likely to have the same pattern

propose new metric to automatically filter generations

subject generated examples to human review, where crowdworkers (optionally) revise for quality and assign a gold label
The subset of ambiguous examples in a dataset leads to improved generalization (Swayamdipta et al., 2020) and has fewer spurious lexical correlations (Gardner et al., 2021).
Stage 1: Collection ("knowing the unknowns")

1) Find seed examples that are valuable for training

Use examples belonging to the most ambiguous 25% of MultiNLI relative to RoBERTa-large (M)

2) For each seed, collect a group of similar examples having the same "reasoning" pattern

Use the $k = 4$ nearest neighbors (that have the same label) in terms of [CLS] token representation in $M$

Captures "reasoning" similarity, rather than semantic or lexical similarity!
automatically **collect** pockets of examples that exemplify valuable reasoning patterns

leverage GPT-3 to **generate** new examples likely to have the same pattern

propose new metric to automatically **filter** generations

subject generated examples to **human review**, where crowdworkers (optionally) revise for quality and assign a gold label
Stage 2: Overgeneration

Given a group of examples, we create a context for GPT-3. Differently from its traditional usage in few-shot settings, we generate examples rather than predict labels. These examples don’t have a gold label!

Write a pair of sentences that have the same relationship as the previous examples.
Examples:
1. \{premise\}
Implication: \{hypothesis\}
5. \{premise\}
Implication: \{hypothesis\}

Template for entailment examples
1. In **six states**, the federal investment represents almost the entire contribution for providing civil legal services to low-income individuals. 
**Implication:** In **44 states**, the federal investment does not represent the entire contribution for providing civil legal services for people of low income levels.

2. But if it's at all possible, plan your visit for the **spring, autumn, or even the winter**, when the big sightseeing destinations are far less crowded. 
**Implication:** This destination is most crowded in the **summer**.

3. **5 percent** of the routes operating at a loss. 
**Implication:** **95 percent** of routes are operating at either profit or break-even.

4. About **10 percent** of households did not 
**Implication:** Roughly **ninety percent** of households did this thing.

5. **5 percent** probability that each part will be defect free. 
**Implication:** Each part has a **95 percent** chance of having a defect.

6. **1 percent** of the seats were vacant. 
**Implication:** **99 percent** of the seats were occupied.
At the time of the Revolution, the old port of Marseille was a great center of shipbuilding and commerce.

**Possibility:** The only place where ships were built was in the old port of Marseille.
1. Dun Laoghaire is the major port on the **south coast**.  
**Contradiction**: Dun Laoghaire is the major port on the **north coast**.

2. Leave the city by its **eastern** Nikanor Gate for a five-minute walk to Hof Argaman (Purple Beach), one of Israel's finest beaches.  
**Contradiction**: Leave the city by its **western** Nikanor Gate for a fifty-five-minute walk to Hof Argaman.

3. **Southwest** of the Invalides is the Ecole Militaire, where officers have trained since the middle of the 18th century.  
**Contradiction**: **North** of the Invalides is the Ecole Militaire, where officers have slept since the early 16th century.

4. Across the courtyard on the **right-hand side** is the chateau's most distinctive feature, the splendid Francois I wing.  
**Contradiction**: The Francois I wing can be seen across the courtyard on the **left-hand side**.

5. To the **south**, in the Sea of Marmara, lie the woods and beaches of the Princes' Islands.  
**Contradiction**: In the **north** is the Sea of Marmara where there are mountains to climb.

6. From the park's **southern entrance**, follow the avenue **south** to the Hotel de Ville.  
**Contradiction**: From the park's **northern entrance**, follow the avenue **north** to the Hotel de Ville.
automatically collect pockets of examples that exemplify valuable reasoning patterns

leverage GPT-3 to generate new examples likely to have the same pattern

propose new metric to automatically filter generations

subject generated examples to human review, where crowdworkers (optionally) revise for quality and assign a gold label
Stage 3: Map-Based Filtering

The ambiguity $\sigma_i$ of a example $(x_i, y_i)$ is defined by the standard deviation in the probability assigned to the correct label $y_i$ across a model’s $E$ epochs of training

$\sigma_i = \sigma \left( \left\{ p_{\theta(e)}(y_i | x_i) \right\}_{e \in E} \right)$

Problem: we don’t have a gold label

Now, given a new unlabeled example $x_i$, how can we estimate its ambiguity without any additional training?

💡 We can save the checkpoints $\theta^{(e)}$ and retroactively compute the predictions on a new example.
Stage 3: Map-Based Filtering

The ambiguity $\sigma_i$ of a example $(x_i, y_i)$ is defined by the standard deviation in the probability assigned to the correct label $y_i$ across a model’s $E$ epochs of training

$$\sigma_i = \max_{y \in \mathcal{Y}} \sigma \left( \left\{ p_{\theta(e)}(y \mid x_i) \right\}_{e \in E} \right)$$

(Swayamdipta et al., 2020)

Now, given a new unlabeled example $x_i$, how can we estimate its ambiguity without any additional training?

- Problem: we don’t have a gold label
- Solution: take the “worst case” over all labels

estimated max variability

We can save the checkpoints $\theta^{(e)}$ and retroactively compute the predictions on a new example
automatically collect pockets of examples that exemplify valuable reasoning patterns

leverage GPT-3 to generate new examples likely to have the same pattern

propose new metric to automatically filter generations

subject generated examples to human review, where crowdworkers assign a gold label and (optionally) revise for quality
Stage 4: Human review

Each example is reviewed by 2 crowdworkers on AMT

1) **Premise:** He claimed that he had been pressured into giving a false confession.

   **Hypothesis:** He had been pressured into giving a false confession.

   *(Optional) Revise the example below.*

   **Premise:**
   He claimed that he had been pressured into giving a false confession.

   **Hypothesis:**
   He had been pressured into giving a false confession.

   **Given the premise, the hypothesis is...**
   
<table>
<thead>
<tr>
<th>Definitely correct</th>
<th>Maybe correct, maybe not</th>
<th>Definitely incorrect</th>
<th>Discard</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Entailment</em></td>
<td><em>Neutral</em></td>
<td><em>Contradiction</em></td>
<td></td>
</tr>
</tbody>
</table>

   assign a label   OR   discard if it would take a great deal of revision to fix, or it could be perceived as **offensive**
1) Improve the fluency of the text

P: He had no idea that he was the only one in the room.
H: He was the only one in the room,
    he was the only one in the room.
    Entailment

P: There is a slight possibility that, if the same temperature data are used, the temperature of the Earth's surface in 1998 will be lower than the temperature of the Earth's surface now.
H: The Earth's surface in 1998 was lower than the Earth's surface now.
    Neutral

2) Improve the clarity of the relationship

P: As I climbed the mountain, I noticed that the clouds were parting, and the sun was shining through.
H: The sun was shining through the clouds.
    Entailment

P: This will be the first time the king has met the queen in person.
H: The king has met the queen in person before.
    Contradiction
Inherent ambiguities in NLI

1) P: According to the most recent statistics, the rate of violence crime in the United States has dropped by almost half since 1991. 
H: The rate of violent crime has not dropped by half since 1991.

Does “almost half” mean “not half” or “basically half”?

2) P: He’d made it clear that he was not going to play the game.
H: He didn't want to play the game.

Can we assume intention behind actions?

3) P: If you can’t handle the heat, get out of the kitchen.
H: If you can't handle the pressure, get out of the situation.

Is the premise to be interpreted literally or metaphorically?

4) P: As a result of the disaster, the city was rebuilt and it is now one of the most beautiful cities in the world.
H: A disaster made the city better.

Do indirect consequences count? Does “more beautiful” even mean “better”?

5) P: It is a shame that the world has to suffer the pain of such unnecessary war.
H: The world does not have to suffer such pain.

What is the scope of “have to” in the hypothesis?
Dataset Statistics

118,724 examples → keep? (91% revise? (both workers revised)

neither worker discarded

Train: 103,079 examples
Label distribution (E/N/C): 38,609 / 49,053 / 15,418

Test: 5,000 examples
Label distribution (E/N/C): 1,858 / 2,397 / 745
Does training on WANLI improve model robustness?

WANLI leads to better OOD generalization than MNLI across the board, despite being ~4x smaller.
Does training on WANLI improve in-domain performance?

Including WANLI in the training data can improve in-domain test performance too!
Exploration of Artifacts

Compared to MultiNLI, WaNLI has

less information about the label contained in the hypothesis alone

fewer previously known lexical correlations (e.g., “because”, “never”, “nothing”)

less information about the label contained in the semantic similarity between the premise and hypothesis
Takeaways

New approach for the creation of NLP datasets based on LM generation and human labeling & revision

Applied it to create a new dataset for NLI, which we showed leads to more robust models while avoiding known issues in existing NLI datasets

How can we distill human linguistic competence into datasets that models can learn from and be evaluated on?

This work: ask workers to revise and evaluate content, rather than write free-form examples
What’s next?

How should we deal with the inherent ambiguities in NLI examples?

What are other ways of defining valuable examples, and of leveraging generation to create those examples?

How can we leverage the generation + revision idea without relying on an existing large-scale dataset?