Abstract

We present a large scale unified natural language inference (NLI) dataset for providing insight into how well sentence representations capture distinct types of reasoning. We generate a large-scale NLI dataset by recasting 11 existing datasets from 7 different semantic tasks. We use our dataset of approximately half a million context-hypothesis pairs to test how well sentence encoders capture distinct semantic phenomena that are necessary for general language understanding. Some phenomena that we consider are event factuality, named entity recognition, figurative language, gendered anaphora resolution, sentiment analysis, extending prior work that included semantic roles and frame semantic parsing. Our dataset will be available at https://www.decomp.net, to grow over time as additional resources are recast.

1 Introduction

How well do sentence representation learning models capture semantic phenomena necessary for general natural language understanding? Can models trained to make inferences based on natural language perform intelligent reasoning? Do competitive models for popular natural language inference1 (NLI) datasets carry-out human-like reasoning? For example, can these models determine whether an event occurred, correctly differentiate between figurative and literal language, or accurately identify and categorize named entities?

We introduce a large-scale NLI dataset to help answer these questions. Our dataset contains context-hypothesis pairs automatically labeled by recasting semantic annotations from structured prediction tasks into NLI sentence pairs. We extend various prior works on challenge datasets, e.g. Zhang et al. (2017) and White et al. (2017). We recast annotations from a total of 11 datasets across 7 NLP tasks into labeled NLI examples. The tasks include event factuality, named entity recognition, gendered anaphora resolution, sentiment analysis, relationship extraction, pun detection, and lexico-syntactic inference. Currently, our dataset contains approximately half a million labeled sentence-pairs. Table 1 includes a sample of NLI pairs that test specific types of reasoning.

Our experiments demonstrate how this dataset can be used to probe how well a model captures different types of semantic reasoning necessary for general natural language understanding (NLU). We use a NLI model with access to just hypothesis sentences as a baseline, since it has recently been shown to be a strong baseline on many NLI datasets and may be used to detect statistical irregularities within class labels (Poliak et al., 2018b).

2 Motivation & Background

2.1 Why automatic recasting?

Compared to other methods used to construct NLI datasets, we believe that recasting, i.e. leveraging existing annotations to automatically create labeled NLI pairs, can 1) help determine whether an NLU model performs distinct types of reasoning, 2) limit biases in NLI data, and 3) generate labeled NLI examples efficiently at large scales.

NLU insights

Recasting allows us to create targeted subsets of our NLI dataset that can be used as a benchmark

1The task of determining whether a human would likely infer a textual hypothesis from a context, or premise (Dagan et al., 2006, 2013).

2By design, such a model does not reason about any relationship between a pair of texts.

3This site will be continuously updated as we recast more datasets.
Table 1: Example sentence pairs for the different semantic phenomena. The ✓ column indicates that the context entails the hypothesis and the ✗ column lists examples where the context does not entail the hypothesis. The first and second line in each cell respectively represents the context and hypothesis. Examples here are slightly modified for aesthetic reasons.

<table>
<thead>
<tr>
<th>Semantic Phenomena</th>
<th>✓</th>
<th>✗</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event Factuality</td>
<td>Find him before he finds the dog food.</td>
<td>I’ll need to ponder.</td>
</tr>
<tr>
<td></td>
<td>The finding did not happen.</td>
<td>The pondering happened.</td>
</tr>
<tr>
<td>Named Entity Recognition</td>
<td>She travels to Liberia on Sunday.</td>
<td>She travels to Liberia on Sunday.</td>
</tr>
<tr>
<td></td>
<td>Liberia is a location</td>
<td>Liberia is a day of the week.</td>
</tr>
<tr>
<td>Gendered Anaphora</td>
<td>The chef came out to apologize to the guest who was unhappy with his/her dinner.</td>
<td>The paramedic performed CPR on the passenger even though he/she was already dead. The paramedic was already dead.</td>
</tr>
<tr>
<td>Lexicosyntactic Inference</td>
<td>a particular person remembered to do a particular thing</td>
<td>a particular person wished to have a particular thing</td>
</tr>
<tr>
<td></td>
<td>that person did that thing</td>
<td>that person had that thing</td>
</tr>
<tr>
<td>Relation Extraction</td>
<td>Ward went with Ledger to their native Perth.</td>
<td>Stefan had visited his son in Bulgaria.</td>
</tr>
<tr>
<td></td>
<td>Ward was born in Perth.</td>
<td>Stefan was born in Bulgaria.</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td>When asked about the restaurant, Liam said, “The food came out at a good pace.”</td>
<td>When asked about the movie, Eyana said, “The character developments lacked depth.”</td>
</tr>
<tr>
<td></td>
<td>Liam liked the restaurant</td>
<td>Eyana liked the movie.</td>
</tr>
<tr>
<td>Pun</td>
<td>Kim heard masks have no face value</td>
<td>Tod heard that thrift is better than an annuity</td>
</tr>
<tr>
<td></td>
<td>Kim heard a pun</td>
<td>Tod heard a pun</td>
</tr>
<tr>
<td>Thematic Roles</td>
<td>I searched in the cave for treasure.</td>
<td>John informed me of the situation.</td>
</tr>
<tr>
<td></td>
<td>The cave was searched</td>
<td>John received information.</td>
</tr>
</tbody>
</table>

to test a model’s ability to perform distinct types of reasoning. For example, by recasting annotations focused on labeling puns, we can determine if a model may be able to understand whether a certain text is figurative or literal.

Popular NLI datasets, e.g. Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) and its successor Multi-NLI (Williams et al., 2017), were created by eliciting humans to generate hypotheses. Crowd-source workers were tasked with writing one sentence that is entailed, one that is neutral, and one that is contradicted by a given context extracted from the Flickr30k corpus (Young et al., 2014). Although these datasets are widely used to evaluate sentence representations, a high accuracy is not indicative of what types of reasoning NLI models perform. Workers were free to create any type of hypothesis sentence as long as it was neutral, entailed, or contradicted with respect to the context. These datasets cannot be used to determine how well an NLI model captures paraphrastic inference, complex anaphora resolution (White et al., 2017), or compositionality (Pavlick and Callison-Burch, 2016; Dasgupta et al., 2018), which are among the many desired capabilities of language understanding systems.

**Limit biases**

Recent studies demonstrate that many NLI datasets contain significant biases. For example, sentences in SNLI contain racial, gender, and age-related stereotypes (Rudinger et al., 2017). Also, statistical irregularities, and annotation artifacts, within class labels allow a model with access only to the hypotheses to outperform the majority baseline by over 2x (Poliak et al., 2018b; Gururangan et al., 2018). Class label biases may be attributed to the human-elicited protocol used to generate hypotheses. The human-elicitation protocol can also cause inconsistencies in how different NLI examples deal with some linguistic phenomena, e.g. questions and person.\(^4\)

We limit some biases by not relying on humans to generate hypotheses. Recast NLI datasets may still contain some biases, e.g. non-uniform distributions over NLI labels caused by the distribu-

\(^4\)Appendix A further discusses this issue.
tion of labels in the original dataset that we recast.\footnote{This does not invalidate the recast dataset. For example, in a corpus annotated with part-of-speech tags, the distribution of labels for the word “the” will likely peak at the Det tag.} Experimental results using Poliak et al. (2018b)’s hypothesis-only model indicate to what degree the recast datasets retain biases that may be present in the original semantic datasets that we recast.

**Large-scale NLI costs**

Generating NLI datasets from scratch is costly. Humans must be paid to generate or label natural language text. This linearly scales costs as the amount of generated NLI-pairs increases. Existing annotations for a wide array of semantic NLP tasks are freely available. By leveraging such datasets, we can automatically generate and label NLI pairs at little cost.

### 2.2 Why these semantic phenomena?

A long term goal is to develop NLU systems that can achieve human levels of understanding and reasoning. Investigating how different architectures and training corpora can help a system perform human-level general NLU is an important step in this direction. Our dataset contains recast NLI pairs that are easily understandable by humans and can be used as to evaluate different sentence encoders and NLU systems. These semantic phenomena cover distinct types of reasoning that an NLU system may often encounter in the wild. While higher performance on these benchmarks might not be conclusive proof of a system achieving human-level reasoning, a system that performs poorly should not be viewed as performing human-level NLU. We argue that these semantic phenomena, e.g. event factuality, figurative language, lexical entailment and presupposition, play integral roles in NLU, and acknowledge that understanding other semantic phenomena are also integral to NLU.

### 3 Recasting Semantic Phenomena

We recast 7 semantic phenomena from a total of 11 datasets. Here we describe efforts to recast event factuality, named entity recognition, figurative language, sentiment analysis, and relation extraction into labeled NLI examples. We are currently recasting more semantic phenomena from datasets, like thematic roles in VerbNet (Schuler, 2005), and will release them online\footnote{https://www.decomp.net} as they become available. These 7 semantic phenomena represent distinct types of reasoning and inferences that play a key role in general NLU.

Many of the recasting methods rely on simple templates that do not include nuances and variances typical of natural language, e.g. ((3a)). This allows us to specifically test how sentence representations capture distinct types of reasoning. When recasting, we preserve each dataset’s train/development/test split. If a dataset does not contain such a split, we randomly sample pairs to create a split with a 80:10:10 ratio. Table 2 reports statistics about each already recast NLI dataset.

#### 3.1 Event Factuality (EF)

Event factuality prediction is the task of determining whether an event described in text actually occurred. Determining whether an event occurred enables accurate inferences based on the event (Rudinger et al., 2018b). For example, consider the following sentences:

(1)  
- a. She walked a beagle  
- b. She walked a dog  
- c. She walked a brown beagle

If the walking occurred, (1a) entails (1b) but not (1c). If we negate the action in sentences (1a), (1b), and (1c) to respectively become:

(2)  
- a. She did not walk a beagle  
- b. She did not walk a dog  
- c. She did not walk a brown beagle

the new hypothesis (2c) is now entailed by the context (2a) while (2b) is not. In order to make monotonic inferences based on text that describes an event, a system must be able to latently predict whether the event happened. Incorporating factuality has been shown to improve natural language inference (Saurf and Pustejovsky, 2007).

We recast event factuality annotations from three existing datasets: UW (Lee et al., 2015), MEANTIME (Minard et al., 2016), and the JHU Universal Decompositional Semantics (Decomp It Happened v2 (Rudinger et al., 2018b).\footnote{The last dataset provides a larger coverage of White et al. (2016)’s event factuality annotations on top of the English version of the Universal Dependencies treebank (Silveira et al., 2014).} Context sentences are simple sentences from the three...
datasets. We use the templates in (3a) and (3b) as hypotheses such that each context is paired with two hypotheses.

(3)  
   a. The Event happened  
   b. The Event did not happen

If the predicate denoting the Event was annotated as having happened in the factuality dataset, the context paired with (3a) is labeled as entailed and the context paired with (3b) is labeled as not-entailed. If the Event was labeled as not having happened, we swap the labels.

### 3.2 Named Entity Recognition

Distinct types of entities have different properties and relational objects (Prince, 1978). Relational objects can be used to help infer facts from a given context. For example, if a system can detect that an entity is a name of a nation, then that entity likely has a leader, a language, and a culture (Prince, 1978; Van Durme, 2010). When classifying NLI pairs, a model can determine if an object mentioned in the hypothesis can be a relational object typically associated with the type of entity described in the context. NER tags can also be directly used to determine if a hypothesis is not entailed by a context. If entities in contexts and hypotheses do not share NER tags, e.g. PERSON or ORGANIZATION, then the hypothesis may be unlikely to be entailed by the context. Pipeline approaches to NLI successfully incorporate this intuition (Castillo and Alemany, 2008; Sammons et al., 2009; Pakray et al., 2010).

When recasting NER annotations, we preserve the original sentences as premises and create hypotheses using the following template:

(4) NP is a Label

To generate an entailed hypothesis we replace Label with the correct NER label of the NP and to generate a not-entailed hypothesis we replace it with an incorrect label. When creating not-entailed hypotheses, we choose the incorrect label from the prior distribution of NER tags for the given phrase. This prevents us from adding additional annotation artifacts besides any class-label statistical irregularities present in the original NER data.

---

We apply this recasting strategy on the Groningen Meaning Bank (Bos et al., 2017b) and the ConLL-2003 NER Shared Task (Tjong Kim Sang and De Meulder, 2003). In total, we generate about 320K labeled context-hypothesis pairs from NER annotations.

### 3.3 Gendered Anaphora Resolution

The ability to perform pronoun resolution is essential to language understanding, in many cases requiring common-sense reasoning about the world, as demonstrated by the Winograd Schema Challenge (Levesque et al., 2012). White et al. (2017) show that this task can be directly recast as a natural language inference problem by transforming Winograd schemas into NLI sentence pairs; in the transformation, the pronoun in question is simply replaced with one of the two possible antecedents.

Using a formula similar to Winograd schemas, Rudinger et al. (2018a) introduce Winogender schemas, minimal sentence pairs that differ only by pronoun gender.9 Using this adapted pronoun resolution task, Rudinger et al. (2018a) demonstrate the presence of systematic gender bias in coreference resolution systems. Here we recast Winogender schemas as an NLI task, introducing a potential method of detecting gender bias in NLI systems or sentence embeddings. In recasting, the context is the original, unmodified Winogender sentence; the hypothesis is a short, manually constructed sentence that entails one of the two possible pronoun resolutions.

### 3.4 Lexicosyntactic Inference

While many inferences in natural language are triggered by lexical items alone, there exist pervasive inferences that arise from interactions between lexical items and their syntactic contexts. This is particularly apparent among propositional attitude verbs – e.g. think, want, know – which display complex distributional profiles (White and Rawlins, 2016). For instance, the verb remember can take both finite clausal complements and infinitival clausal complements.

(5)  
   a. Jo didn’t remember that she ate.  
   b. Jo didn’t remember to eat.

This small change in the syntactic structure gives...
Table 2: Statistics summarizing the recast datasets. ‘Annotations’ column refers to the original annotation that was recast, the ‘Combined’ row refers to the combination of our recast datasets. The ‘Dataset’ column indicates the datasets that were recast, and the ‘# Pairs’ column reports how many labeled NLI pairs were extracted from the corresponding dataset (numbers in parentheses indicate the exact number of NLI pairs). We include the number of examples in Multi-NLI and SNLI in order to compare the scale of our Unified NLI Framework to popular NLI datasets.

<table>
<thead>
<tr>
<th>Annotations</th>
<th>Dataset</th>
<th># pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event Factuality</td>
<td>Universal Dependencies (Rudinger et al., 2018b)</td>
<td>42K (41,888)</td>
</tr>
<tr>
<td></td>
<td>MeanTime (Minard et al., 2016)</td>
<td>.7K (738)</td>
</tr>
<tr>
<td></td>
<td>UW (Lee et al., 2015)</td>
<td>5K (5,094)</td>
</tr>
<tr>
<td>Named Entity Recognition</td>
<td>Groningen (Bos et al., 2017a)</td>
<td>260K (261,406)</td>
</tr>
<tr>
<td></td>
<td>CoNLL (Tjong Kim Sang and De Meulder, 2003)</td>
<td>60K (59,970)</td>
</tr>
<tr>
<td>Gendered Anaphora</td>
<td>Winogender (Rudinger et al., 2018a)</td>
<td>.4K (464)</td>
</tr>
<tr>
<td>Lexicosyntactic Inference</td>
<td>MegaAttitude (White and Rawlins, 2016, 2018)</td>
<td>11K (11,814)</td>
</tr>
<tr>
<td>Relationship Extraction</td>
<td>FACC1 (Gabrilovich et al., 2013)</td>
<td>30K</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td>(Kotzias et al., 2015)</td>
<td>6K</td>
</tr>
<tr>
<td>Puns</td>
<td>(Yang et al., 2015)</td>
<td>7K (7,706)</td>
</tr>
<tr>
<td></td>
<td>SemEval 2017 Task 7 (Miller et al., 2017)</td>
<td>8K (8,054)</td>
</tr>
<tr>
<td>Combined</td>
<td>Unified NLI Framework</td>
<td>433K (433,134)</td>
</tr>
<tr>
<td></td>
<td>SNLI (Bowman et al., 2015)</td>
<td>570K</td>
</tr>
<tr>
<td></td>
<td>Multi-NLI (Williams et al., 2017)</td>
<td>433K</td>
</tr>
</tbody>
</table>

rise to large changes in the inferences that are licensed: (5a) presupposes that Jo ate while (5b) entails that Jo didn’t eat.

We recast data collected by White and Rawlins (2018), which is relevant to these sorts of lexicosyntactic interactions. White and Rawlins selected verbs from the MegaAttitude data (White and Rawlins, 2016) based on their acceptability in the [NP _ that S] and [NP was _ed that S] frames (a rating of 4 out of 7 or better). They then asked annotators to answer questions of the form in (6) using three possible responses: yes, maybe or maybe not, and no (cp. Karttunen et al., 2014).

(6) a. Someone [knew, didn’t know] that a particular thing happened.
   b. Did that thing happen?

We use the same procedure to annotate sentences containing verbs that take various types of infinitival complement: [NP _ for NP to VP], [NP _ to VP], [NP _ NP to VP], and [NP was _ed to VP].

To recast these annotations, we call the sentences like (6a) the context sentence and assign each to the majority class – yes, maybe or maybe not, no – across 10 different annotators, after applying an ordinal model-based normalization to those annotators’ responses. We then pair each context sentence with three hypothesis sentences.

(7) a. That thing happened.
   b. That thing may or may not have happened.
   c. That thing didn’t happen.

If the sentence is assigned to yes, the pairing of the context sentence with (7a) is assigned entailed and the other pairings are assigned not-entailed; if the sentence is assigned maybe or maybe not, the pairing of the context sentence with (7b) is assigned entailed and the other pairings are assigned not-entailed; and if the sentence is assigned no, the pairing of the context sentence with (7c) is assigned entailed and the other pairings are assigned not-entailed. Training, development, and test split labels are randomly assigned by context sentence to every pair that context sentence appears in.
3.5 Relation extraction (KG relations)

The goal of the relation extraction task is to infer the real-world relationships between pairs of entities from natural language text. The task is “grounded” in the sense that the input is natural language text and the output is (entity1, relation, entity2) tuples defined in the schema of some knowledge base. Relation extraction requires a system to understand many different surface forms which entail the same underlying relation, and to distinguish those from surface forms which involve the same entities but do not entail the relation of interest. For example, (8a) is entailed by both (8b) and (8c) but not by (8d).

(8)  a. Name was born in Place
    b. Name is from Place
    c. Name (b. Year, Place)
    d. Name visited Place

Natural language surface forms are often used in relation extraction in a weak-supervision setting (Mintz et al., 2009; Hoffmann et al., 2011; Riedel et al., 2013). That is, if entity1 and entity2 are known to be related by relation, it is assumed that every sentence observed which mentions both entity1 and entity2 is assumed to be a realization of relation: i.e. (8d) would (falsely) be taken as evidence of the birthPlace relation. Therefore, in order to ensure high-quality in our recast data, we have human annotators vet each generated context/hypothesis pair. Our procedure for recasting the data is described below.

Unlike the previous recasting methods, here we first generate hypotheses and then corresponding contexts. To generate our hypotheses, we begin with entity-relation triples extracted from DBPedia infoboxes: e.g. ⟨Barack Obama, birthPlace, Hawaii⟩. In order to recast each tuple to be a natural language string, we use the canonical forms of entities given by DBPedia, and generate a gloss for the relation as follows: for a given relation, we enumerate all pairs of entities appearing with that relation, and compute, on aggregate, the natural language predicate most likely to relate entity1 to entity2 according to a large OpenIE database built using the OpenIE 5.0 software (Christensen et al., 2011; Pal et al., 2016; Saha et al., 2017). This translates a tuple such as that above to a string such as “Barack Obama was born in Hawaii”.

For each hypothesis generated as described above, we create a number of contexts, both true and false, as follows. We begin with the FACC1 corpus (Gabrilovich et al., 2013) which contains natural language sentences from ClueWeb in which entities have been automatically linked to disambiguated Freebase entities, when possible. Then, given a tuple (entity1, relation, entity2), we find every sentence which contains both entity1 and entity2. Since many of these sentences are false positives (see Example 8d), we have human annotators vet each context/hypothesis pair, using the ordinal entailment scale described in Zhang et al. (2017). In total, we generate 30K labeled context-hypothesis pairs using this process, 38% of which are entailments and 62% of which are non-entailments.

3.6 Subjectivity

Some of the previously discussed semantic phenomena deal with objective information – did an event occur or what type of entities does a specific name represent. Subjective information is often expressed differently than objective information (Wiebe et al., 2005), making it important to use different tests to probe whether a NLU system understands language the expresses subjective information. Therefore, we are interested in determining whether general NLU models capture ‘subjective clues’ that can help identify and understand emotions, opinions, and sentiment within a subjective text (Wilson et al., 2006), as opposed to differentiating between subjective and objective information (Yu and Hatzivassiloglou, 2003; Riloff et al., 2003).

We recast a sentiment analysis dataset since the task is the “expression of subjectivity as either a positive or negative opinion” (Taboada, 2016). We extract sentences from product, movie, and restaurant reviews labeled with positive or negative sentiment towards the reviewed item (Kotzias et al., 2015). NLI contexts (9a) and hypotheses (9b), (9c) are generated using the following templates:

(9)  a. When asked about Item, Name said Review
    b. Name liked the Item

12 This results in a dataset with hypotheses that appear multiple times, similar to Khot et al. (2018)’s SciTail.
c. *Name* did not like the *Item*

where *Item* ∈ \{product, movie, restaurant\}. We sample names from a distribution based on data provided by the U.S. Social Security Agency.\(^{13}\) If the original sentence was labeled as containing positive sentiment, the (9a)-(9b) pair is labeled as entailed and the (9a)-(9c) pair is labeled as not-entailed; otherwise, the labels are swapped. We use this method to generate 6K examples. A model performing well on this recast dataset may indicate that it may be able to sufficiently identify polarity in subjective text.

### 3.7 Figurative Language

Figurative language demonstrates natural language’s expressiveness and wide variations. Understanding and recognizing figurative language “entails[s] cognitive capabilities to abstract and meta-represent meanings beyond physical words” (Reyes et al., 2012). Puns are prime examples of figurative language that may perplex general NLU systems as they are one of the more regular uses of linguistic ambiguity (Binsted, 1996) and rely on a wide-range of phonetic, morphological, syntactic, and semantic ambiguity (Pepicello and Green, 1984; Binsted, 1996; Bekinschtein et al., 2011). To understand a pun, a NLU system may need to know that two words (or phrases) with different spelling may sound the same, associate misspelled words with multiple real words, or understand how a polysemous word’s senses may simultaneously be used in context.\(^{14}\)

We recast puns from Yang et al. (2015)\(^{15}\) and Miller et al. (2017)\(^{16}\) using the templates to generate contexts (10a) and hypotheses (10b), (10c). Each context appears in two labeled NLI pairs.

\begin{equation}
\begin{aligned}
(10) \quad \text{a.} & \quad \text{Name heard that } \text{Pun} \\
\text{b.} & \quad \text{Name heard a pun} \\
\text{c.} & \quad \text{Name did not hear a pun}
\end{aligned}
\end{equation}

For the contexts, we insert the original pun into *Pun*. We use the same method as previously described to insert names into the template. If the original sentence was labeled as containing a pun, the (10a)-(10b) pair is labeled as entailed and the (10a)-(10c) pair is labeled as not-entailed, otherwise we swap the NLI labels. In total, we generate roughly 15K labeled pairs.

### 4 Experiments

Our experiments demonstrate how these recast datasets can be used to evaluate how well models capture different types of semantic reasoning necessary for general language understanding. We also include results from a model with access to just hypotheses as a strong baseline. This may indicate whether the recast datasets retain statistical irregularities with-in class labels from the original, task-specific annotations.

#### 4.1 Models

For demonstrating how well an NLI model performs these fine-grained types of reasoning, we use InferSent (Conneau et al., 2017). Our hypothesis-only model is a modified version of InferSent that only accesses hypotheses (Poliak et al., 2018b). This determines whether a model can perform well on each recast NLI dataset when only considering the context. We use the same training settings and hyper-parameters as described in Conneau et al. (2017) and Poliak et al. (2018b).

#### 4.2 Results

Table 3 reports the models’ accuracies across the recast NLI datasets. We use each model under three different scenarios. In the first two scenarios, we train and test a model separately on each recast dataset. The model’s parameters are either initialized randomly or with parameters extracted from a model pre-trained on Multi-NLI (pre-train, update). In the third scenario, we test a model trained only on Multi-NLI and test it on each semantic phenomena (pre-train, fixed). Because of the small size of the recast Gendered Anaphora Resolution dataset, we do not use it for training.

Evaluating NLI models When evaluating NLI models, our baseline is the maximum of the accuracies of the hypothesis-only model and always predicting the majority class label. The hypothesis-only model can demonstrate how...
likely a NLI label applies to a hypothesis, regardless of its context. Our strong baseline indicates how well each recast dataset tests a model’s ability to perform each specific type of reasoning when determining whether a context entails a hypothesis. Our results suggest that InferSent might capture specific semantic phenomena better than others. InferSent seems to learn the most about determining if a described event occurred since the difference between its accuracy and baseline is largest on the recast Event Factuality dataset compared to the recast annotations.

The high hypothesis-only accuracy on the recast NER dataset may demonstrate that the hypothesis-only model is able to detect that the distribution of class labels for a given word may be peaky. Based on this, we may need to consider different methods to recast NER annotations into labeled NLI examples, or limit the size of the dataset’s training partition.

### Models trained on Multi-NLI
Williams et al. (2017) argue that Multi-NLI makes “it possible to evaluate systems on nearly the full complexity of the language.” However, how well does Multi-NLI test a models capability to understand distinct semantic phenomena often occurring in language? We posit that if a model, trained on and performing well on Multi-NLI, does not perform well on our recast datasets, then it is possible that Multi-NLI might not evaluate a model’s ability to understand the “full complexity” of language as thought.

When trained on Multi-NLI, our InferSent model achieves an accuracy of 70.22% on Multi-NLI. When we test the model on the recast datasets, we see significant drops as it performs below every majority baseline. This might suggest that Multi-NLI does not evaluate whether sentence representations capture these distinct semantic phenomena. Multi-NLI’s train set contains text from five domains: fiction, government, slate, telephone, and travel. We would expect the fiction section, and especially its humor subset, to contain some figurative language that might be similar to puns, and the travel guides (and possibly telephone conversations) to contain text related to sentiment.

### Initializing with pre-trained parameters
We would like to know whether initializing models with parameters trained on Multi-NLI improve scores. On three of the recast datasets, there is minimal difference when using a pre-trained model. We notice a large improvement on the Puns and Sentiment recast datasets.

### 5 Related Work

#### Targeted Tests for Natural Language Understanding
We follow a long line of work focused on building datasets to test how well NLU systems perform distinct types of semantic reasoning. FraCaS uses a limited number of sentence-pairs to test whether systems understand semantic phenomena such as generalized quantifiers, temporal references, and (nominal) anaphora (Cooper et al., 1996). FraCas cannot be used to train neural models – it includes just roughly 300 high-quality instances manually created by linguists. MacCartney (2009) created the FraCaS textual in-
ference test suite by automatically “convert[ing] each FraCaS question into a declarative hypothesis.” Levesque et al. (2012)’s Winograd Schema Challenge force a model two choose between two possible answers for a question based on a sentence describing an event. Additionally, public benchmarks exist that test whether RTE models are unable to handle adjective-noun composition (Pavlick and Callison-Burch, 2016), other types of composition (Dasgupta et al., 2018), paraphrastic inference, anaphora resolution, and semantic proto-roles (White et al., 2017).

Exploring what linguistic phenomena neural models learn Many tests have been used to probe how well neural models learn different linguistic phenomena. Linzen et al. (2016) use “number agreement in English subject-verb dependencies” to show that LSTMs learn about syntax-sensitive dependencies. In addition to syntax (Shi et al., 2016), researchers have used other labeling tasks to investigate whether neural machine translation (NMT) models learn other types of linguistic phenomena (Belinkov et al., 2017b,a; Dalvi et al., 2017; Marvin and Koehn, 2018). Recently, Poliak et al. (2018a) recently argued for using White et al. (2017)’s recast NLI datasets to investigate whether NMT encoders capture semantic phenomena.

6 Conclusion

We have described how we recast a wide range of semantic phenomena from many NLP datasets into labeled NLI sentence pairs. These examples serve as a unified NLI framework that may help diagnose whether NLU models capture and perform distinct types of reasoning. Our experiments demonstrate how to use this framework can as a NLU benchmark. Our dataset is actively growing as we continue recasting more datasets into labeled NLI examples. The dataset, along with baselines, trained models, and a leader board, will be publicly available at https://www.decomp.net.

Acknowledgements

We thank Diyi Yang for help with the PunsOfTheDay dataset and Tongfei Chen for suggestions regarding designing the figures and tables.

References


of the sixth international conference on Knowledge capture, pages 113–120. ACM.


Dimitrios Kotzias, Misha Denil, Nando De Freitas, and Padhraic Smyth. 2015. From group to individual labels using deep features. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 597–606. ACM.


A Linguistic considerations

Williams et al. (2017) define labels for context-hypothesis pairs in SNLI and Multi-NLI as:

“one which is necessarily true or appropriate in the same situations as the premise (to be paired with the premise under the label entailment), one which is necessarily false or inappropriate whenever the premise is true (contradiction), and one where neither condition applies (neutral).”

Depending on different linguistic phenomena in a context, there are multiple equally-valid, yet contradictory ways to understand what the same situations actually are. Multi-NLI is inconsistent in understanding the same situations when different linguistic structures make the same situations ambiguous. Table 4 shows some examples...
A.1 Person

Consider the following sentences where (11a) is the context and (11b) and (11c) are two hypotheses entailed by the context:

(11) a. She walked a beagle
    b. A dog got exercise
    c. A woman exercised
d. You walked a beagle

When we replace She in context (11a) with You, (11b) would be entailed by the new context (11d) but what about (11c)? From the single sentence context, it is unclear whether the person being addressed is the woman who walked the beagle or someone else.

There are at least three potential solutions: The conservative approach ignores all premise sentences that are not in the third person. This solution prevents issues that arise in Multi-NLI but decreases the size of our dataset. The second solution, which we refer to as the observer approach, is to change personal pronouns when generating hypotheses. Under this design, (11c) would now become (12a).

(12) a. I exercised.

The third approach can be considered the speaker’s perspective. Here, we define our hypotheses labeled as entailed as being new utterances that any agent who provided the natural language context would be willing to commit to as well. Here, (11c) is changed to (13a).

(13) a. You exercised.

A.2 Questions

Imagine changing the context-sentence in (11a) to be now posed as the question (14a).

(14) a. She walked a beagle?

In NLI, is a question a rhetorical statement invalidating the plausibility of a scenario ((15a)’s speaker claiming there is no way she walked the beagle) or a way to clarify understanding about a situation that occurred ((15b)’s speaker wants to know whether she walked a beagle or a different type of dog)?

(15) a. She walked a beagle?
    b. She walked a beagle?

The context-hypotheses pair’s entailment label depends on how one understands this question, whereby (11c) is entailed by (15b) but not (15a).

The most obvious solution is to remove questions in the set of context sentences. A downside to this solution is that potentially many NLI contexts will be disregarded. Other solutions may include forcing valid hypotheses to answer the question, or forcing valid hypotheses to be other questions that are subsets (entailed) or supersets (not-entailed) or unrelated (not-entailed) with respect to the question in the context.