Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation

Adam Poliak[1] **Aparajita Haldar**[1,2] **Rachel Rudinger**[1] **J. Edward Hu**[1]
Ellie Pavlick[3] **Aaron Steven White**[4] **Benjamin Van Durme**[1]
[1]Johns Hopkins University, [2] BITS Pilani, Goa Campus, India
[3]Brown University, [4]University of Rochester

Abstract

We present a large scale collection of diverse natural language inference (NLI) datasets that help provide insight into how well a sentence representation captures distinct types of reasoning. The collection results from recasting 13 existing datasets from 7 semantic phenomena into a common NLI structure, resulting in over half a million labeled context-hypothesis pairs in total. We refer to our collection as the DNC: ***D***iverse ***N***atural Language Inference ***C***ollection. The DNC is available online at `http://www.decomp.net`, and will grow over time as additional resources are recast and added from novel sources.

1 Introduction

A plethora of new natural language inference (NLI)[1] datasets has been created in recent years (Bowman et al., 2015; Williams et al., 2017; Lai et al., 2017; Khot et al., 2018). However, these datasets do not provide clear insight into what type of reasoning or inference a model may be performing. For example, these datasets cannot be used to evaluate whether competitive NLI models can determine if an event occurred, correctly differentiate between figurative and literal language, or accurately identify and categorize named entities. Consequently, these datasets cannot answer how well sentence representation learning models capture distinct semantic phenomena necessary for general natural language understanding (NLU).

To answer these questions, we introduce the **D**iverse **N**LI **C**ollection (DNC), a large-scale NLI dataset that tests a model's ability to perform diverse types of reasoning. DNC is a collection of NLI problems, each requiring a model to perform a unique type of reasoning. Each NLI dataset contains labeled context-hypothesis pairs that we recast from semantic annotations for specific structured prediction tasks. We extend various prior works on challenge NLI datasets (Zhang et al., 2017), and define recasting as leveraging existing datasets to create NLI examples (Glickman, 2006; White et al., 2017). We recast annotations from a total of 13 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 lexicosyntactic inference. Currently, the DNC contains over half a million labeled examples. Table 1 includes NLI pairs that test specific types of reasoning.

Event Factuality	▶ Find him before he finds the dog food The finding did not happen	✓
	▶ I'll need to ponder The pondering happened	✗
Relation Extraction	▶ Ward joined Tom in their native Perth Ward was born in Perth	✓
	▶ Stefan had visited his son in Bulgaria Stefan was born in Bulgaria	✗
Puns	▶ Kim heard masks have no face value Kim heard a pun	✓
	▶ Tod heard that thrift is better than annuity Tod heard a pun	✗

Table 1: Example sentence pairs for different semantic phenomena. ▶ indicates the line is a context and the following line is its corresponding hypothesis. ✓ and ✗ respectively indicate that the context entails, or does not entail the hypothesis. Appendix A includes more recast examples.

Using a *hypothesis-only* NLI model, with access to just hypothesis sentences, as a strong baseline (Tsuchiya, 2018; Gururangan et al., 2018; Poliak et al., 2018b), our experiments demonstrate how DNC can be used to probe a model's ability to capture different types of semantic reasoning

[1]The task of determining if a hypothesis would likely be inferred from a context, or premise; also known as Recognizing Textual Entailment (RTE) (Dagan et al., 2006, 2013).

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 67–81
Brussels, Belgium, October 31 - November 4, 2018.

necessary for general NLU. In short, this work answers a recent plea to the community to test "more kinds of inference" than in previous challenge sets (Chatzikyriakidis et al., 2017).

2 Motivation & Background

Compared to eliciting NLI datasets directly, i.e. asking humans to author contexts and/or hypothesis sentences, recasting can 1) help determine whether an NLU model performs distinct types of reasoning; 2) limit types of biases observed in previous NLI data; and 3) generate examples cheaply, potentially at large scales.

NLU Insights Popular NLI datasets, e.g. Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) and its successor MultiNLI (Williams et al., 2017), were created by eliciting hypotheses from humans. Crowd-source workers were tasked with writing one sentence each that is entailed, neutral, and contradicted by a caption extracted from the Flickr30k corpus (Young et al., 2014). Although these datasets are widely used to train and 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 for each context and label. Such datasets cannot be used to determine how well an NLI model captures many desired capabilities of language understanding systems, e.g. paraphrastic inference, complex anaphora resolution (White et al., 2017), or compositionality (Pavlick and Callison-Burch, 2016; Dasgupta et al., 2018). By converting prior annotation of a specific phenomenon into NLI examples, recasting allows us to create a diverse NLI benchmark that tests a model's ability to perform distinct types of reasoning.

Limit Biases Studies indicate that many NLI datasets contain significant biases. Examples in the early Pascal RTE datasets could be correctly predicted based on syntax alone (Vanderwende and Dolan, 2006; Vanderwende et al., 2006). Statistical irregularities, and annotation artifacts, within class labels allow a hypothesis-only model to significantly outperform the majority baseline on at least six recent NLI datasets (Poliak et al., 2018b). Class label biases may be attributed to the human-elicited protocol. Moreover, examples in such NLI datasets may contain racial and gendered stereotypes (Rudinger et al., 2017).

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 distribution of labels in the original dataset that we recast.[2] Experimental results using Poliak et al. (2018b)'s hypothesis-only model indicate to what degree the recast datasets retain some biases that may be present in the original semantic datasets.

NLI Examples at Large-scale 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 existing semantic annotations already invested in by the community we can generate and label NLI pairs at little cost and create large NLI datasets to train data hungry models.

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. DNC contains recast NLI pairs that are easily understandable by humans and can be used 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 does poorly should not be viewed as performing human-level NLU. We argue that these semantic phenomena play integral roles in NLU. There exist more semantic phenomena integral to NLU (Allen, 1995) and we plan to include them in future versions of the DNC.

Previous Recast NLI Example sentences in RTE1 (Dagan et al., 2006) were extracted from MT, IE, and QA datasets, with the process referred to as 'recasting' in the thesis by Glickman (2006). NLU problems were reframed under the NLI framework and candidate sentence pairs were extracted from existing NLP datasets and then labeled under NLI (Dagan et al., 2006). Years later, this term was independently used by White et al.

[2] In a corpus with part-of-speech tags, the distribution of labels for the word "the" will likely peak at the `Det` tag.

(2017), who proposed to "leverage existing large-scale semantic annotation collections as a source of targeted textual inference examples." The term 'recasting' was limited to automatically converting existing semantic annotations into labeled NLI examples without manual intervention. We adopt the broader definition of 'recasting' since our NLI examples were automatically or manually generated from prior NLU datasets.

Applied Framework versus Inference Probing Traditionally, NLI has not been viewed as a downstream, applied NLP task.[3] Instead, the community has often used it as "a generic evaluation framework" to compare models for distinct downstream tasks (Dagan et al., 2006) or to determine whether a model performs distinct types of reasoning (Cooper et al., 1996). These two different evaluation goals may affect which datasets are recast. We target both goals as we recast applied tasks and linguistically focused phenomena.

3 Recasting Semantic Phenomena

We describe efforts to recast 7 semantic phenomena from a total of 13 datasets into labeled NLI examples. Many of the recasting methods rely on simple templates that do not include nuances and variances typical of natural language. This allows us to specifically test how sentence representations capture distinct types of reasoning. When recasting, we preserve each dataset's train/dev/test split. If a dataset does not contain such a split, we create a random split with roughly a 80:10:10 ratio. Table 2 reports statistics about each recast dataset.

Event Factuality (EF) Event factuality prediction is the task of determining whether an event described in text occurred. Determining whether an event occurred enables accurate inferences, e.g. monotonic inferences, based on the event (Rudinger et al., 2018b).[4] Incorporating factuality has been shown to improve NLI (Sauri and Pustejovsky, 2007).

We recast event factuality annotations from UW (Lee et al., 2015), MEANTIME (Minard et al., 2016), and Decomp (Rudinger et al., 2018b). We use sentences from original datasets as contexts and templates (1a) and (1b) as hypotheses.[5]

(1) 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 (1a) is labeled as ENTAILED and the same context paired with (1b) is labeled as NOT-ENTAILED. Otherwise, we swap the labels.

Named Entity Recognition (NER) Distinct types of entities have different properties and relational objects (Prince, 1978) that can 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 likely to not be entailed by a context, such as when entities in contexts and hypotheses do not share NER tags (Castillo and Alemany, 2008; Sammons et al., 2009; Pakray et al., 2010).

Given a sentence annotated with NER tags, we recast the annotations by preserving the original sentences as contexts and creating hypotheses using the template "*NP* is a *Label*."[6] For ENTAILED hypotheses we replace *Label* with the correct NER label of the *NP*; for NOT-ENTAILED hypotheses, we choose an incorrect label from the prior distribution of NER tags for the given phrase. This prevents us from adding additional biases besides any class-label statistical irregularities present in the original data. We apply this procedure on the Gronigen Meaning Bank (Bos et al., 2017) and the ConLL-2003 Shared Task (Tjong Kim Sang and De Meulder, 2003).

Gendered Anaphora Resolution (GAR) The ability to perform pronoun resolution is essential to language understanding, in many cases requiring common-sense reasoning about the world (Levesque et al., 2012). White et al. (2017) show that this task can be directly recast as an NLI problem by transforming Winograd schemas into NLI sentence pairs.

Using a similar formula Rudinger et al. (2018a) introduce *Winogender schemas*, minimal sentence pairs that differ only by pronoun gender. With this

[3] This changed as large NLI datasets have recently been used to train, or pre-train, models to perform NLI, or other tasks (Conneau et al., 2017; Pasunuru and Bansal, 2017).

[4] Appendix B.1 provides an example.

[5] We replace *Event* with the event described in the context.

[6] We ensure grammatical hypotheses by appropriately conjugating "is a" when needed.

Sem. Phenomena	Dataset	# pairs	Automated
Event Factuality	Decomp (Rudinger et al., 2018b)	42K (41,888)	✓
	UW (Lee et al., 2015)	5K (5,094)	✓
	MeanTime (Minard et al., 2016)	.7K (738)	✓
Named Entity Recognition	Groningen (Bos et al., 2017)	260K (261,406)	✓
	CoNLL (Tjong Kim Sang and De Meulder, 2003)	60K (59,970)	✓
Gendered Anaphora	Winogender (Rudinger et al., 2018a)	.4K (464)	✗
Lexicosyntactic Inference	VerbCorner (Hartshorne et al., 2013)	135K (138, 648)	✓
	MegaVeridicality (White and Rawlins, 2018)	11K (11,814)	✓
	VerbNet (Schuler, 2005)	2K (1, 759)	✓✗
Puns	(Yang et al., 2015)	9K (9,492)	✓
	SemEval 2017 Task 7 (Miller et al., 2017)	8K (8, 054)	✓
Relationship Extraction	FACC1 (Gabrilovich et al., 2013)	25K (25,132)	✓✗
Sentiment Analysis	(Kotzias et al., 2015)	6K (6,000)	✓
Combined	Diverse NLI Collection (DNC)	570K (570,459)	
—	SNLI (Bowman et al., 2015)	570K	
—	Multi-NLI (Williams et al., 2017)	433K	

Table 2: Statistics summarizing the recast datasets. The first column refers to the original annotation that was recast, the 'Combined' row refers to the combination of our recast datasets. The second column indicates the datasets that were recast, and the 3rd column reports how many labeled NLI pairs were extracted from the corresponding dataset. The last column indicates whether the recasting method was fully-automatic without human involvement (✓), manual (✗), or used a semi-automatic method that included human intervention (✓✗). The Multi-NLI and SNLI numbers contextualize the scale of our dataset.

adapted pronoun resolution task, they demonstrate the presence of systematic gender bias in coreference resolution systems. 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 having a correct (ENTAILED) or incorrect (NOT-ENTAILED) pronoun resolution.

Lexicosyntactic Inference (Lex) 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.

(2) a. Jo didn't **remember** *that she ate*
b. Jo didn't **remember** *to eat*

This small change in the syntactic structure gives rise to large changes in the inferences that are licensed: (2a) presupposes that *Jo ate* while (2b) entails that *Jo didn't eat*. We recast data from three datasets that are relevant to these sorts of lexicosyntactic interactions.

Lex #1: MegaVeridicality (MV) White and Rawlins (2018) build the MegaVeridicality dataset by selecting verbs from the MegaAttitude dataset (White and Rawlins, 2016) based on their grammatical acceptability in the [NP _ that S] and [NP was _ed that S] frames.[7] They then asked annotators to answer questions of the form in (3) using three possible responses: *yes*, *maybe or maybe not*, and *no* (Karttunen et al., 2014).

(3) 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].[8]

To recast these annotations, we assign the context sentences like (3a) to the majority class – *yes*, *maybe or maybe not*, *no* – across 10 different annotators, after applying an ordinal model-based normalization to their responses. We then pair each context sentence with three hypotheses.

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

[7]NP is always instantiated by *someone*; and S is always instantiated by *a particular thing happened.*

[8]NP is always instantiated by either *someone*, *a particular person*, or *a particular thing*; and VP is always instantiated by *happen*, *do a particular thing*, or *have a particular thing*.

If annotated *yes*, *maybe or maybe not*, or *no*, the pair (3a)-(4a), (3a)-(4b), or (3a)-(4c) is respectively assigned ENTAILED and the other pairings are assigned NOT-ENTAILED; train/dev/test split labels are randomly assigned to every pair that context sentence appears in.

Lex #2: Recasting VerbNet (VN) We create additional lexicosyntactic NLI examples from VerbNet (Schuler, 2005). VerbNet contains classes of verbs that each can have multiple frames. Each frame contains a mapping from syntactic arguments to thematic roles, which are used as arguments in Neo-Davidsonian first-order logical predicates (5b) that describe the frame's semantics. Each frame additionally contains an example sentence (5a) that we use as our NLI context and we create templates (5c) from the most frequent semantic predicates to generate hypotheses (5d).

(5) a. Michael swatted the fly
    b. cause(*E*, *Agent*)
    c. *Agent* caused the *E*
    d. Michael caused the swatting

We use the Berkeley Parser (Petrov et al., 2006) to match tokens in an example sentence with the thematic roles and then fill in the templates with the matched tokens (5d). We also decompose multi-argument predicates into unary predicates to increase the number of hypotheses we generate. On average, each context is paired with 4.5 hypotheses. We generate NOT-ENTAILED hypotheses by filling in templates with incorrect thematic roles. [9] We partition the recast NLI examples into train/development/test splits such that all example sentences from a VerbNet class (which we use a NLI hypothesis) appear in only one partition of our dataset. In turn, the recast VerbNet dataset's partition is not exactly 80:10:10.

Lex #3: Recasting VerbCorner (VC) The third dataset testing lexicosyntactic inference that we recast is VerbCorner (VC) (Hartshorne et al., 2013). VC decomposes VerbNet predicates into simple semantic properties and "elicit[s] reliable semantic judgments corresponding to VerbNet predicates" via crowd-sourcing. The semantic judgments focus on movement, physical contact, application of force, change of physical or mental state, and valence, all of which "may be central organizing principles for a human's ... conceptualization of the world." (Hartshorne et al., 2013).

Each sentence in VC is judged based on the decomposed semantic properties. We convert each semantic property into declarative statements[10] to create hypotheses and pair them with the original sentences which we preserve as contexts. The NLI pair is ENTAILED or NOT-ENTAILED depending on the given sentence's semantic judgment.

Figurative Language (Puns) Figurative language demonstrates natural language's expressiveness and wide variations. Understanding and recognizing figurative language "entail[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).

We recast puns from Yang et al. (2015) and Miller et al. (2017) using templates to generate contexts (6a) and hypotheses (6b), (6c). We replace *Name* with names sampled from a distribution based on US census data,[11] and *Pun* with the original sentence. If the original sentence was labeled as containing a pun, the (6a)-(6b) pair is labeled as ENTAILED and (6a)-(6c) is labeled as NOT-ENTAILED, otherwise we swap the labels.

(6) a. *Name* heard that *Pun*
    b. *Name* heard a pun
    c. *Name* did not hear a pun

Relation Extraction (RE) The goal of the relation extraction (RE) 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. RE requires a system to understand the many different surface forms which may 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, (7a) is entailed by (7b) and (7c) but not by (7d).

[9] This is similar to Aharon et al. (2010)'s template matching to generate entailment rules from FrameNet (Baker et al., 1998).

[10] We list the declarative statements in Appendix B.2.1.

[11] `http://www.ssa.gov/oact/babynames/names.zip`

(7) a. *Name* was born in *Place*
b. *Name* is from *Place*
c. *Name*, a *Place* native, . . .
d. *Name* visited *Place*

Natural language surface forms are often used in RE 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. (7d) would (falsely) be taken as evidence of the `birthPlace` relation.

Here we first generate hypotheses and then corresponding contexts. To generate hypotheses, we begin with entity-relation triples extracted from DBPedia infoboxes: e.g. ⟨*Barack Obama*, `birthPlace`, *Hawaii*⟩. These relation predicates were extracted directly from Wikipedia infoboxes and are not cleaned. As a result, many relations are redundant with one another (`birthPlace`, `hometown`) and some relations do not correspond to obvious natural language glosses based on the name alone (`demographics1Info`). Thus, we construct a template for each predicate `p` by manually inspecting 1) a sample of entities which are related by `p` 2) a sample of sentences in which those entities co-occur and 3) the most frequent natural language strings which join entities related by `p` according to a OpenIE triple database (Schmitz et al., 2012; Fader et al., 2011) extracted from a large text corpus. We then manually write a simple template (e.g. Mention1 *was born in* Mention2) for `p`, ignoring any unclear relations. In total, we end up with 574 unique relations, expressed by 354 unique templates.

For each such hypothesis generated, we create a number of contexts. 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 (7d), we have human annotators vet each context/hypothesis pair, using the ordinal entailment scale described in Zhang et al. (2017). We include optional binary labels by converting pairs labeled as $1-4$ and 5 to ENTAILED and NOT-ENTAILED respectively.[12] We apply pruning methods (described in Appendix B.4) to combat issues related to noisy, ungrammatical hypotheses and disagreement between multiple annotators.

Subjectivity (Sentiment) 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 (Wiebe et al., 2005), making it important to use other tests to probe whether an NLU system understands language that expresses subjective information. 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).

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 as containing positive or negative sentiment (Kotzias et al., 2015). Contexts (8a) and hypotheses (8b), (8c) are generated using the following templates:

(8) a. When asked about *Item*, *Name* said *Review*
b. *Name* liked the *Item*
c. *Name* did not like the *Item*

Item is replaced with either "product", "movie", or "restaurant", and the *Name* is sampled as previously discussed. If the original sentence contained positive (negative) sentiment, the (8a)-(8b) pair is labeled as ENTAILED (NOT-ENTAILED) and (8a)-(8c) is labeled as NOT-ENTAILED (ENTAILED).

3.1 Noise in Recast Data

Recasting can create noisy NLI examples that may potentially enable a model to achieve a high accuracy by learning dataset specific characteristics that are unrelated to NLU. For example, Poliak et al. (2018a,b) previously noted the association between ungrammaticality and NOT-ENTAILED examples based on how White et al. (2017) recast the FrameNet+ dataset (Pavlick et al., 2015).

[12] Following the label set in SNLI, Zhang et al. (2017) converted pairs labeled with 1 as CONTRADICTION, $2-4$ as NEUTRAL and 5 to ENTAILMENT. Since here we are generally interested in binary classification, we merge the CONTRADICTION and NEUTRAL examples as NOT-ENTAILED.

Model \ Recast Data	NER	EF	RE	Puns	Sentiment	GAR	VC	MV	VN
Majority (MAJ)	50.00	50.00	59.53	50.00	50.00	50.00	50.00	66.67	53.66
No Pre-training									
InferSent	**92.50**	83.07	61.89	60.36	50.00	–	88.60	**85.96**	46.34
Hyp-only	91.48	69.14	64.78	60.36	50.00	–	76.82	77.83	46.34
Pre-trained DNC									
InferSent *(update)*	92.47	**83.86**	74.38	**93.17**	81.00	–	**89.00**	85.62	76.83
InferSent *(fixed)*	92.20	81.07	74.11	87.76	77.33	**50.65**	88.59	83.84	67.68
Hyp-only *(update)*	91.60	71.07	70.57	60.02	46.83	–	76.78	77.83	68.90
Hyp-only *(fixed)*	91.37	69.74	65.97	56.44	48.17	50.00	76.78	77.83	59.15
Pre-trained Multi-NLI									
InferSent *(update)*	92.37	83.03	**76.08**	92.48	**83.50**	–	88.45	85.11	**78.05**
InferSent *(fixed)*	52.99	54.88	66.75	56.04	56.50	**50.65**	45.33	55.92	45.73
Hyp-only *(update)*	91.62	70.64	69.91	60.36	49.33	–	76.82	77.83	68.29
Hyp-only *(fixed)*	52.55	66.33	52.96	60.59	50.00	50.43	41.31	46.28	48.78

Table 3: NLI accuracies on test data. Columns correspond to each semantic phenomena and rows correspond to the model used. Columns are ordered from larger to smaller in size, but the last three (VC, MV, VN) are separated since they fall under lexicosyntactic inference. *(update)* refers to a model that was initialized with pre-trained parameters and then re-trained on the corresponding recast data. *(fixed)* refers to a model that was trained and then evaluated on these data sets. Bold numbers in each column indicate which settings were responsible for the highest accuracy on the specific recast dataset.

In the DNC, most of the noisy examples are in the recast VerbNet and Relation Extraction portions. In recast VerbNet, some examples are noisy because of incorrect subject-verb agreement.[13] Since more noisy examples appeared in the Relation Extraction set, we relied on Amazon Mechanical Turk workers to flag ungrammatical hypotheses in the recast dataset, and we remove NLI pairs with ungrammatical hypotheses.[14]

4 Experiments

Our experiments demonstrate how these recast datasets may be used to evaluate how well models capture different types of semantic reasoning necessary for general language understanding. We also include results from a hypothesis-only model as a strong baseline. This may reveal whether the recast datasets retain statistical irregularities 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). `InferSent` independently encodes a context and hypothesis with a bi-directional LSTM and combines the sentence representations by concatenating the individual sentence representations, their element-wise subtraction and product. The combined representation is then fed into a MLP with a single hidden layer. The hypothesis-only model is a modified version of `InferSent` that only accesses hypotheses (Poliak et al., 2018b). We report experimental details in Appendix C.

4.2 Results

Table 3 reports the models' accuracies across the recast NLI datasets. Even though we categorize VerbNet, MegaVeridicality, and VerbCorner as lexicosyntatic inference, we train and evaluate models separately on these three datasets because we use different strategies to individually recast them. When evaluating NLI models, our baseline is the maximum between the accuracies of the hypothesis-only model and the majority class label (MAJ). In six of the eight recast datasets that we use to train our models the hypothesis-only model outperforms MAJ. The two datasets where the hypothesis-only model does not outperform MAJ are Sentiment and VN, each of which contain less than $10K$ examples.[15] We do not train on GAR because of its small size.

Our results suggest that `InferSent`, when not pre-trained on any other data, might capture specific semantic phenomena better than other seman-

[13] "Her *teeth was* cared for" or "*Floss were* used".

[14] See Appendix B.4 for details.

[15] This is similar to Poliak et al. (2018b)'s results where a hypothesis-only model did not outperform MAJ on datasets with $\leq$ 10K examples.

tic phenomena. InferSent seems to learn the most about determining if an event occurred, since the difference between its accuracy and that of the hypothesis-only baseline (+13.93) is largest on the recast EF dataset compared to the other recast annotations. The model seems to similarly learn to perform (or detect) the type of lexicosyntactic inference present in VC and MV. Interestingly, the hypothesis-only model outperforms InferSent on the recast RE.

Hypothesis Only Baseline The hypothesis-only model can demonstrate how likely it is that an NLI label applies to a hypothesis, regardless of its context and indicates how well each recast dataset tests a model's ability to perform each specific type of reasoning when performing NLI. 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. For example, *Hong Kong* appears 130 times in the training set and is always labeled as a location. Based on this, in future work we may consider different methods to recast NER annotations into labeled NLI examples, or limit the dataset's training size.

Pre-training models on DNC We would like to know whether initializing models with pre-trained parameters improves scores. We notice that when we pre-train our models on DNC, for the larger datasets, a pre-trained model does not seem to significantly outperform randomly initializing the parameters. For the smaller datasets, specifically Puns, Sentiment and VN, a pre-trained model significantly outperforms random initialization.[16]

We are also interested to know whether fine-tuning these pre-trained models on each category (*update*) improves a model's ability to perform well on the category compared to keeping the pre-trained models' parameters static (*fixed*). Across all of the recast datasets, updating the pre-trained model's parameters during training improves InferSent's accuracies more than keeping the model's parameters fixed. When updating a model pre-trained on the entire DNC, we see the largest improvements on VN (+9.15).

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 model's capability to understand the diverse semantic phenomena captured in DNC? We posit that if a model, trained on and performing well on Multi-NLI, does not perform well on our recast datasets, then Multi-NLI might not evaluate a model's ability to understand the "full complexity" of language as argued.[17]

When trained on Multi-NLI, our InferSent model achieves an accuracy of 70.22% on (matched) Multi-NLI.[18] When we test the model on the recast datasets (without updating the parameters), we see significant drops.[19] On the datasets testing a model's lexicosyntactic inference capabilities, the model performs below the majority class baseline. On the NER, EF, and Puns datasets its performs below the hypothesis-only baseline. We also notice that on three of the datasets (EF, Puns, and VN), the fixed hypothesis-only model outperforms the fixed InferSent model.

These results might suggest that Multi-NLI does not evaluate whether sentence representations capture these distinct semantic phenomena. This is a bit surprising for some of the recast phenomena. We would expect Multi-NLI's fiction section (especially its humor subset) in the training set 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.

Pre-training on DNC or Multi-NLI? Initializing a model with parameters pre-trained on DNC or Multi-NLI often outperforms random initialization.[20] Is it better to pre-train on DNC or Multi-NLI? On five of the recast datasets, using a model pre-trained on DNC outperforms a model pre-trained on Multi-NLI. The results are flipped on the two datasets focused on downstream tasks (Sentiment and RE) and MV. However, the differences between pre-training on the DNC or Multi-NLI are small. From this, it is unclear whether pre-training on DNC is better than Multi-NLI.

Size of Pre-trained DNC Data We randomly sample 10K and 20K examples from each

[16] By 32.81, 31.00, and 30.83 points respectively.

[17] We treat Multi-NLI's NEUTRAL and CONTRADICTION labels as equivalent to the DNC's NOT-ENTAILED label.

[18] Although this is about 10 points below SoTA, we believe that the pre-trained model performs well enough to evaluate whether Multi-NLI tests a model's capability to understand the diverse semantic phenomena in the DNC.

[19] InferSent (*pre-trained, fixed*) in Table 3.

[20] Pre-training does not improve accuracies on NER or MV.

datasets' training set to investigate what happens if we train our models on a subsample of each training set instead of the entire DNC. Although we noticed a slight decrease across each recast test set, the decrease was not significant. We leave this investigating for a future thorough study.

5 Related Work

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 different linguistic phenomena (Belinkov et al., 2017a,b; Dalvi et al., 2017; Marvin and Koehn, 2018). Recently, Poliak et al. (2018a) used recast NLI datasets to investigate semantics captured by NMT encoders.

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, e.g. 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 inference test suite* by automatically "convert[ing] each FraCaS question into a declarative hypothesis." Levesque et al. (2012)'s Winograd Schema Challenge forces a model to choose between two possible answers for a question based on a sentence describing an event.

Recent benchmarks test whether NLI models 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). Concurrently, Conneau et al. (2018)'s benchmark can be used to probe whether sentence representations capture many linguistic properties. It includes syntactic and surface form tests but does not focus on as a wide range of semantic phenomena as in the DNC. Glockner et al. (2018) introduce a modified version of SNLI to test how well NLI models perform when requiring lexical and world knowledge.

Wang et al. (2018)'s GLUE dataset is intended to evaluate and potentially train a sentence representation to perform well across different NLP tasks. This continues an aspect of the initial RTE collection, designed to be representative of downstream tasks like QA, MT, and IR (Dagan et al., 2010). While GLUE is therefore concerned with applied tasks, DNC, as well as Naik et al. (2018)'s NLI *stress* tests, is concerned with *probing* the capabilities of NLU models to capture explicitly distinguished aspects of meaning. While one may conjecture that the latter is needed to be "solved" to eventually "solve" the former, it may be that these goals only partially overlap. Some NLP researchers might focus on probing for semantic phenomena in sentence representations while others may be more interested in developing single sentence representations that can help models perform well on a wide array of downstream tasks.

6 Conclusion

We described how we recast a wide range of semantic phenomena from many NLP datasets into labeled NLI sentence pairs. These examples serve as a diverse NLI framework that may help diagnose whether NLU models capture and perform distinct types of reasoning. Our experiments demonstrate how to use this framework as an NLU benchmark. The DNC is actively growing as we continue recasting more datasets into labeled NLI examples. We encourage dataset creators to recast their datasets in NLI and invite them to add their recast datasets into the DNC. The collection, along with baselines and trained models are available online at `http://www.decomp.net`.

Acknowledgements

We thank Diyi Yang for help with the PunsOfTheDay dataset, the JSALT "Sentence Representation" team for insightful discussions, and three anonymous reviewers for feedback. This work was supported by the JHU HLTCOE, DARPA LORELEI and AIDA, NSF-BCS (1748969/1749025), and NSF-GRFP (1232825). 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

Roni Ben Aharon, Idan Szpektor, and Ido Dagan. 2010. Generating entailment rules from framenet. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 241–246. Association for Computational Linguistics.

James Allen. 1995. *Natural language understanding*. Pearson.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Tristan A Bekinschtein, Matthew H Davis, Jennifer M Rodd, and Adrian M Owen. 2011. Why clowns taste funny: the relationship between humor and semantic ambiguity. *Journal of Neuroscience*, 31(26):9665–9671.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872. Association for Computational Linguistics.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Kim Binsted. 1996. *Machine humour: An implemented model of puns*. Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland.

Johan Bos, Valerio Basile, Kilian Evang, Noortje J Venhuizen, and Johannes Bjerva. 2017. The groningen meaning bank. In *Handbook of Linguistic Annotation*, pages 463–496. Springer.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Julio Javier Castillo and Laura Alonso Alemany. 2008. An approach using named entities for recognizing textual entailment. In *Notebook Papers of the Text Analysis Conference, TAC Workshop*.

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*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Alexis Conneau, GermÃČÂąn Kruszewski, Guillaume Lample, LoÃČÂƒc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. 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, et al. 1996. Using the framework. Technical report.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches–erratum. *Natural Language Engineering*, 16(1):105–105.

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 tectual entailment*, pages 177–190. Springer.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, and Stephan Vogel. 2017. Understanding and improving morphological learning in the neural machine translation decoder. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 142–151, Taipei, Taiwan. Asian Federation of Natural Language Processing.

I. Dasgupta, D. Guo, A. Stuhlmüller, S. J. Gershman, and N. D. Goodman. 2018. Evaluating Compositionality in Sentence Embeddings. *ArXiv e-prints*.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1535–1545. Association for Computational Linguistics.

Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. 2013. Facc1: Freebase annotation of clueweb corpora, version 1 (release date 2013-06-26, format version 1, correction level 0). *Note: http://lemurproject. org/clueweb09/FACC1/Cited by*, 5.

Oren Glickman. 2006. *Applied textual entailment*. Ph.D. thesis, Bar Ilan University.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. 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 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Joshua K Hartshorne, Claire Bonial, and Martha Palmer. 2013. The verbcorner project: Toward an empirically-based semantic decomposition of verbs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1438–1442.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 541–550, Portland, Oregon, USA. Association for Computational Linguistics.

Lauri Karttunen, Stanley Peters, Annie Zaenen, and Cleo Condoravdi. 2014. The Chameleon-like Nature of Evaluative Adjectives. In *Empirical Issues in Syntax and Semantics 10*, pages 233–250. CSSP-CNRS.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *AAAI*.

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.

Alice Lai, Yonatan Bisk, and Julia Hockenmaier. 2017. Natural language inference from multiple premises. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 100–109, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1648.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, pages 552–561. AAAI Press.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 142–150. Association for Computational Linguistics.

Bill MacCartney. 2009. *Natural language inference*. Ph.D. thesis, Stanford University.

Rebecca Marvin and Philipp Koehn. 2018. Exploring Word Sense Disambiguation Abilities of Neural Machine Translation Systems. In *Proceedings of the 13th Conference of The Association for Machine Translation in the Americas (Volume 1: Research Track*, pages 125–131.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172. ACM.

Tristan Miller and Iryna Gurevych. 2015. Automatic disambiguation of english puns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 719–729.

Tristan Miller, Christian Hempelmann, and Iryna Gurevych. 2017. Semeval-2017 task 7: Detection and interpretation of english puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 58–68, Vancouver, Canada. Association for Computational Linguistics.

Tristan Miller and Mladen Turković. 2016. Towards the automatic detection and identification of english puns. *The European Journal of Humour Research*, 4(1):59–75.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begona Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. Meantime, the newsreader multilingual event and time corpus. In *Language Resources and Evaluation Conference (LREC)*.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Partha Pakray, Santanu Pal, Soujanya Poria, Sivaji Bandyopadhyay, and Alexander F Gelbukh. 2010. Ju_cse_tac: Textual entailment recognition system at tac rte-6. In *TAC Workshop*.

Ramakanth Pasunuru and Mohit Bansal. 2017. Multi-task video captioning with video and entailment generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1273–1283.

Ellie Pavlick and Chris Callison-Burch. 2016. 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. Association for Computational Linguistics.

Ellie Pavlick, Travis Wolfe, Pushpendre Rastogi, Chris Callison-Burch, Mark Dredze, and Benjamin Van Durme. 2015. Framenet+: Fast paraphrastic tripling of framenet. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 408–413, Beijing, China. Association for Computational Linguistics.

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

William J Pepicello and Thomas A Green. 1984. *Language of riddles: new perspectives*. The Ohio State University Press.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 433–440. 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, Louisiana. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018b. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Ellen F Prince. 1978. On the function of existential presupposition in discourse. In *Papers from the... Regional Meeting. Chicago Ling. Soc. Chicago, Ill.*, volume 14, pages 362–376.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–84, Atlanta, Georgia. Association for Computational Linguistics.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018a. Gender bias in coreference resolution. 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 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018b. Neural models of factuality. 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 731–744, New Orleans, Louisiana. Association for Computational Linguistics.

Mark Sammons, VG Vinod Vydiswaran, Tim Vieira, Nikhil Johri, Ming-Wei Chang, Dan Goldwasser, Vivek Srikumar, Gourab Kundu, Yuancheng Tu, Kevin Small, et al. 2009. Relation alignment for textual entailment recognition. In *TAC Workshop*.

Roser Sauri and James Pustejovsky. 2007. Determining modality and factuality for text entailment. In *Semantic Computing, 2007. ICSC 2007. International Conference on*, pages 509–516. IEEE.

Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, et al. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534. Association for Computational Linguistics.

Karin Kipper Schuler. 2005. Verbnet: A broad-coverage, comprehensive verb lexicon.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural mt learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Maite Taboada. 2016. Sentiment analysis: an overview from linguistics. *Annual Review of Linguistics*, 2:325–347.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *11th International Conference on Language Resources and Evaluation (LREC2018)*.

Benjamin Van Durme. 2010. *Extracting Implicit Knowledge from Text*. Ph.D. thesis, University of Rochester, Rochester, NY 14627.

Lucy Vanderwende and William B Dolan. 2006. What syntax can contribute in the entailment task. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 205–216. Springer.

Lucy Vanderwende, Arul Menezes, and Rion Snow. 2006. Microsoft research at rte-2: Syntactic contributions in the entailment task: an implementation. In *Second PASCAL Challenges Workshop*.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

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. 2016. A computational model of s-selection. In *Semantics and linguistic theory*, volume 26, pages 641–663.

Aaron Steven White and Kyle Rawlins. 2018. The role of veridicality and factivity in clause selection. In *Proceedings of the 48th Annual Meeting of the North East Linguistic Society*, page to appear, Amherst, MA. GLSA Publications.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. 2006. Recognizing strong and weak opinion clauses. *Computational intelligence*, 22(2):73–99.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association of Computational Linguistics*, 5(1):379–395.

A More Recast NLI Examples

Table 4 includes examples from all of the recast NLI datasets. We include one ENTAILED and one NOT-ENTAILED example from each dataset that tests a distinct type of reasoning.

B Recasting Semantic Phenomena

Here we add secondary information about the original datasets and our recasting efforts.

B.1 Event Factuality

We demonstrate how determining whether an event occurred can enable accurate inferences based on the event. Consider the following sentences:

(9) a. She walked a beagle
    b. She walked a dog
    c. She walked a brown beagle

If the *walking* occurred, (9a) entails (9b) but not (9c). If we negate the action in sentences (9a), (9b), and (9c) to respectively become:

(10) 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 (10c) is now entailed by the context (10a) while (10b) is not.

B.2 Lexicosyntactic Inference

B.2.1 VerbCorner

When recasting VerbCorner, we use the following templates for hypotheses, assigning them as ENTAILED and NOT-ENTAILED based on the positive or negative answers to the annotation task questions about the context sentence.

(11) a. Someone {moved/did not move} from their location
     b. Something touched another thing / Nothing touched anything else
     c. Someone or something {applied/did not apply} force onto something
     d. Someone or something {changed/did not change} physically
     e. Someone {changed/did not change} their thoughts, feelings, or beliefs
     f. Something {good/neutral/bad} happened

B.3 Figurative Language

Puns in Yang et al. (2015) were originally extracted from `punsoftheday.com`, and sentences without puns came from newswire and proverbs. The sentences are labeled as containing a pun or not. Puns in Miller et al. (2017) were sampled from prior pun detection datasets (Miller and Gurevych, 2015; Miller and Turković, 2016) and includes new examples generated from scratch for the shared task; the original labels denote whether the sentences contain homographic, heterographic, or no pun at all. Here, we are only interested in whether a sentence contains a pun or not instead of discriminating between homographic and heterographic puns.

B.4 Relation Extraction

Since hypotheses were automatically generated from Wikipedia infoboxes, many examples are noisy and ungrammatical. We presented hypotheses (independent of their corresponding contexts) to Mechanical Turk workers and asked them to label each sentence as containing no grammatical error, minor grammatical issues, or major grammatical issues. We removed the $2,056$ NLI examples with hypothesis containing major grammatical issues, resulting in $28,041$ labeled pairs. Interestingly, almost 70% of those examples where labeled between $1-4$, which we view as NOT-ENTAILED. We release the ungrammatical NLI examples as supplementary data.

A second source of noise in the recast relation extraction dataset can be caused by disagreement amongst multiple annotators. Examples in our training and development sets are annotated by a single annotator while we use 3- to 5-way redundancy to annotate the test examples. To guarantee high-quality test examples, we only include examples with 100% inner-annotator agreement. Additionally, we remove the 16 examples labeled with 4 from our NOT-ENTAILED examples in this pruned test set since some of these examples are arguably entailments. Consequently, the test set contains 761 examples, out of the original $3,670$ test examples. Nevertheless, we separately release all $3,670$ test examples and include the original annotations as well, enabling others to consider other methods to collapse the multi-way annotations.

Semantic Phenomena	✓	✗
Event Factuality	I would like to learn how The learning did not happen	I'll not say anything The saying happened
Named Entity Recognition	Ms. Rice said the United States must work intensively Ms. is a person 's title	Afghan officials are welcoming the Netherlands' decision The Netherlands is an event
Gendered Anaphora	The student met with the architect to view her blueprints for inspiration The architect has blueprints	The appraiser told the buyer that he had paid too much for the painting The appraiser had purchased a painting
MegaVeridicality	Someone assumed that a particular thing happened That thing might or might not have happened	A particular person craved to do a particular thing That person did that thing
VerbNet	The Romans destroyed the city The Romans caused the destroying	Andre presented the plaque Andre was transferred
VerbCorner	Molly wheeled Lisa to Rachel Someone moved from their location	Kyle bewildered Mark Someone or something changed physically
Relation Extraction	At least 100,000 Chinese live in Lhasa, outnumbering Tibetans two to one Tibetans live in Lhasa	Tropical storm Humberto is expected to reach the Texas coast tonight Humberto hit Texas
Puns	Jorden heard that my skiing skills are really going downhill Jorden heared a pun	Caiden heard that fretting cares make grey hairs Caiden heared a pun
Sentiment Analysis	When asked about the product, Liam said, "Don't waste your money" Liam did not like the product	When asked about the movie, Angel said, "A bit predictable" Angel liked the movie

Table 4: Example sentence pairs for different semantic phenomena. The ✓ and ✗ columns respectively indicate that the context entails, or does not entail the hypothesis. Each cell's first and second line respectively represent a context and hypothesis.

B.5 Sentiment

Kotzias et al. (2015) compiled examples from previous sources. The movie dataset came from Maas et al. (2011), the Amazon product reviews were released by McAuley and Leskovec (2013) add the restaurant reviews were sourced from the Yelp dataset challenge.[21]

C Experimental Details

In all our experiments, we use pre-computed GloVe embeddings (Pennington et al., 2014) and use the OOV vector for words that do not have a defined embedding. We follow Conneau et al. (2017)'s procedure to train our models. During training, our models are optimized with SGD. Our initial learning rate is 0.1 with a decay rate of 0.99. Our models train for at most 20 epochs and can optionally terminate early when the learning rate is less than 10^{-5}. If the accuracy deceases on the development set in any epoch, the learning rate is divided by 5. As described in Poliak et al. (2018b), our hypothesis-only model feeds the hypotheses' encoded representation directly into the MLP.

[21] http://www.yelp.com/dataset_challenge