Synthetic Dataset for Evaluating Complex Compositional Knowledge for Natural Language Inference

Sushma Anand Akoju, Robert Vacareanu, Haris Riaz, Mihai Surdeanu, Eduardo Blanco

University of Arizona

{sushmaakoju, rvacareanu, hriaz, msurdeanu, eduardoblanco}@arizona.edu

Abstract

We introduce a synthetic dataset called Sentences Involving *Complex* Compositional Knowledge (SICCK) and a novel analysis that investigates the performance of Natural Language Inference (NLI) models to understand compositionality in logic. We produce 1,304 sentence pairs by modifying 15 examples from the SICK dataset (Marelli et al., 2014). To this end, we modify the original texts using a set of phrases – modifiers that correspond to universal quantifiers, existential quantifiers, negation, and other concept modifiers in Natural Logic (NL) (MacCartney, 2009). We use these phrases to modify the subject, verb, and object parts of the premise and hypothesis. Lastly, we annotate these modified texts with the corresponding entailment labels following NL rules. We conduct a preliminary verification of how well the change in the structural and semantic composition is captured by neural NLI models, in both zero-shot and fine-tuned scenarios. We found that the performance of NLI models under the zero-shot setting is poor, especially for modified sentences with negation and existential quantifiers. After fine-tuning this dataset, we observe that models continue to perform poorly over negation, existential and universal modifiers.

1 Introduction

Natural language inference (NLI) has made tremendous progress in recent years, both in terms of datasets, e.g., SNLI (Bowman et al., 2015b), MultiNLI (Williams et al., 2018), Adversarial NLI (Nie et al., 2019), NLI_XY (Rozanova et al., 2021), MonaLog (Hu et al., 2019), and methods (Yang et al., 2020; Lan et al., 2020; Wang et al., 2021b,a; Devlin et al., 2019). However, many of these directions lack explainability, a critical drawback that limits their applicability to critical domains such as medical, legal, or financial. In contrast, Natural Logic (NL) (MacCartney, 2009) provides the necessary explainability through explicit *compositionality* that is driven by several relations that serve as building blocks (Forward Entailment (FE), Reverse Entailment (RE), Negation, Cover, Alternation, Equivalence, and Independence) as well as rules to combine them, which model changes in monotonicity.

In this work, we analyze how well transformer networks trained for NLI understand the atomic reasoning blocks defined in NL, and how well they can compose them to detect changes in monotonicity (Richardson et al., 2020; Joshi et al., 2020). To this end, we create a dataset containing 1304 sentences by modifying 15 premise/hypothesis pairs from the SICK dataset (Marelli et al., 2014). The dataset is generated by modifying the premise and hypothesis sentences selected, as follows:

- We append a series of modifiers to subject/verb/objects in the hypothesis/premise pairs. These modifiers include universal quantifiers (e.g., *every*, *always*), existential quantifiers (e.g., *some*, *at least*), negation, and adverbs/adjectives (e.g., *happy*, *sad*). Table 2 lists the complete set of modifiers used.

- We store the adjusted entailment label for each modifier pair to understand the shift in meaning from word-level changes within sentential contexts. More formally, we used the seven entailment relations as defined in (MacCartney, 2009). These labels were generated manually for each example by following monotonicity calculus and natural logic. For example, consider the premise: *an old man is sitting in a field* and the hypothesis: *a man is sitting in a field*, with the original SICK label: Forward Entailment. After adding the universal quantifier *every* to the aforementioned SICK example, the modified premise: *an old man is sitting in a field* and the original hypothesis: **every** *man is sitting in a field* are annotated

with the adjusted label: Reverse Entailment.

Using this dataset, we analyzed the capacity of three different NLI methods to correctly capture the change in entailment given the modified texts. In particular, the contributions of this work are as follows:

1. We propose a mechanism to generate synthetic data for NLI that enforces compositionality in reasoning. Following this mechanism, we produce 1,304 examples from 15 SICK (Marelli et al., 2014) premise, hypothesis sentence pairs by modifying the sentences for subject, verb, and object respectively with a series of modifiers. The resulting dataset is freely available at https://github.com/clulab/releases/tree/sushma/acl2023-nlrse-sicck.

2. We define specific annotation guidelines based on monotonicity calculus and natural logic (MacCartney, 2009) for annotating the modified premise and hypothesis sentences in the dataset above. The resulting labels are included in the dataset.

3. We conducted an analysis to understand how well these structural and compositional changes are captured by neural NLI models, in both zero-shot and fine-tuned scenarios. Our analysis indicates that NLI models perform poorly over negation and several types of quantifiers. Fine-tuned NLI models do not show significant improvement in learning about compositional changes when compared to their zero-shot equivalent models over our dataset. This suggests that compositionality in reasoning remains a challenge for neural models of language.

2 Related Work

Natural Logic (NL) is a formal reasoning approach that makes use of syntactic structure and semantic properties of lexical items to understand compositionally (MacCartney, 2009).

Logical reasoning is a known challenge for neural NLI models (Ravichander et al., 2019). In particular, NLI models struggle to understand quantifiers, which is highlighted by the fact that these models do not generalize well over quantifier-driven inference tasks (Haruta et al., 2020). The monotonicity calculus over quantifiers with token-level polarity has been explored using the CCG parser over the SICK dataset to generate a synthetic dataset that considers compositional data augmentation (Marelli et al., 2014) and monotonicity calculus (Hu et al., 2019). Other recent research focused on language structures to highlight the importance of compositionality, i.e., the premise and hypothesis differ only in the order of the words, or the presence of antonyms, synonyms, or negation (Dasgupta et al., 2018). Having such data augmentation can help move closer to the compositional encoding of the language (Dasgupta et al., 2018). Our work extends this direction: our dataset captures both phrasal changes (e.g., synonyms, hypernyms), which we inherit from the SICK dataset (Marelli et al., 2014), as well as multiple types of modifiers that are critical for NLI such as universal, existential, negation, and adjectives/adverbs.

The FraCas test suite (Cooper et al., 1996) contains 346 examples that explore aspects of natural logic applied to NLI (MacCartney, 2009). The HELP dataset (Yanaka et al., 2019b) modifies phrases in premise/hypothesis sentences based on monotonicity reasoning from combinatorial categorical grammar (Steedman and Baldridge, 2011) and semantic tagging (Abzianidze and Bos, 2017). As mentioned above, our work is complementary to such datasets, as we cover other types of text modifications. The MED dataset (Yanaka et al., 2019a) is another manually-labeled dataset where hypotheses were also modified by the human labelers given the monotonicity information for the premises. Similarly, we manually labeled NLI information, but our work focuses mainly on compositional information in a sentential context.

Enhancing the dataset with data augmentation is another recent method to test the generalizability of NLI models (Jha et al., 2020). Lexical entailment acquired from the distributional behavior of word pairs (Geffet and Dagan, 2005) led to the subsequent work of (Bowman et al., 2015a), who produced a 3-way classification task for NLI dataset that serves as a benchmark for evaluating natural language understanding. Using Natural Logic as a means to learn and reason about the semantic and lexical relations is a common method used to improve the reasoning capabilities of the NLI models (Bowman et al., 2015c).

The NLI_XY dataset (Rozanova et al., 2021) conducts structural investigation over the

transformer-based NLI models. In particular, the authors investigate how monotonicity (upwards or downwards) changes when the premises and hypotheses are modified through the insertion of hypernym/hyponym phrases. This work is complementary to ours: while they focus on monotonicity in lexicalization (e.g., changing from a hypernym to a hyponym), we focus on changes in monotonicity due to explicit modifiers applied on top of such lexical modifications.

The MonaLog system (Hu et al., 2019) introduces a simple yet explainable NLI method that relies on a simplified Natural Logic implementation. The proposed method operates by implementing monotonicity calculus over CCG syntactic trees using "a small inventory of monotonicity facts about quantifiers, lexical items and token-level polarity." Despite its simplicity, the authors report excellent performance on the SICK dataset. More closely related to our work, they use MonaLog to generate additional training data for NLI from the generated proofs.

3 Dataset

We introduce a synthetic dataset to facilitate the analysis of compositionality in logic. The dataset contains 1,304 sentences that were created by modifying 15 examples from the SICK dataset (Marelli et al., 2014) with a variety of modifiers. To this end, we used a set of phrases that correspond to universal quantifiers, existential quantifiers, negation, and other concept modifiers in Natural Logic (NL) (MacCartney, 2009). These modifiers were applied to syntactic constructs in both premise and hypothesis and the entailment labels are adjusted, as detailed below.

3.1 Overview

At a high level, our dataset creation followed the following steps:

1. We start with 15 seed pairs of premise and hypothesis sentences from SICK. Table 1 shows the seed sentence pairs.

2. We syntactically analyze these sentences to understand their subject-verb-object (SVO) structures. Each of the SVO elements is then modified using a subset of the applicable modifiers listed in Table 2. This process is detailed in Section 3.2.

3. Lastly, we re-annotate the entailment labels for the modified sentences, using the seven entailment relations defined in (MacCartney, 2009): Forward Entailment (FE), Reverse Entailment (RE), Negation (Neg) (or Contradiction), Alternation, Cover, Independence (Neutral) and Equivalence (Equiv). This step is detailed in Section 3.3. The labels are described in Table 3.

3.2 Sentence Modification Strategy

For each premise and hypothesis sentence pair, we modified individual subject, verb, and object phrases with the following approach:

1. To modify *subjects*, we used the Berkeley Neural Parser to extract the left-most noun phrases (NPs). We then append the applicable modifiers from Table 2. In particular, we used universal quantifiers, existential quantifiers, negations, and adjectives.

2. To modify *verbs*, we used the Berkeley Neural parser to extract the rightmost verb phrases (VPs) from the parse tree and appended the applicable modifiers. Verbs were modified using universal quantifiers (*always*, *never*), negations (*not*, *never*), and adverbs (*abnormally*, *elegantly*).

3. To detect *objects*, we used the syntactic dependency parser of (Vacareanu et al., 2020) to identify noun phrases attached to the main verb. Similarly to the subject modifications, these objects were modified using universal quantifiers, existential quantifiers, negations, and adjectives.

After modifying each of the premises and hypotheses sentences, we generate multiple new data points as follows: $f(P_i, H_i, m, SVO) = P_i^{'}, H_i^{'}$ where $m \in M$: all modifiers; SVO : subject/verb/object phrases for either one of the parts of the sentence; and P_i, H_i are premise and hypothesis from sentence pairs $S_i \in S$ where S is the set of 15 examples from SICK. Lastly, f is the function that modifies a given premise and hypothesis that follows one of the modification strategies described above. We generate the following pairs of combinations of the premise, and hypothesis sentences: $(P_i^{'}, H_i), (P_i, H_i^{'}), (P_i^{'}, H_i^{'})$. We repeat this process to modify each of the relevant sentence phrases, as well as a couple of combinations:

Premise	Hypothesis	SICK label
an **old man** is sitting in a field	a **man** is sitting in a field	Entailment
A boy is standing in the **cold water**	A boy is standing in the **water**	Entailment
Two children are **hanging on a large branch**	Two children are **climbing a tree**	Entailment
A **boy** is hitting a baseball	A **child** is hitting a baseball	Entailment
Two dogs are playing by a **tree**	Two dogs are playing by a **plant**	Entailment
A player is **throwing the ball**	**Two teams are competing in a football match**	Neutral
A man is **sitting** in a field	A man is **running** in a field	Neutral
Two dogs are **playing** by a tree	Two dogs are **sleeping** by a tree	Neutral
A girl with a black bag is on a crowded train	**A cramped black train is on the bag of a girl**	Neutral
A blond girl is **riding the waves**	A blond girl is **looking at the waves**	Neutral
The **turtle** is following the **fish**	The **fish** is following the **turtle**	Contradiction
A man is jumping into an **empty** pool	A man is jumping into a **full** pool	Contradiction
A deer **is jumping** over a fence	A deer **isn't jumping** over the fence	Contradiction
A child is **hitting** a baseball	A child is **missing** a baseball	Contradiction
A classroom is **full of students**	A classroom is **empty**	Contradiction

Table 1: 15 premise/hypothesis sentence pairs from the SICK dataset (Marelli et al., 2014) and corresponding NLI labels that form the seed of our dataset. The bold text highlights the lexically-driven compositional change in the premise and hypothesis sentences.

Modifier Type	Modifiers
Universal quantifiers	*every, always, never, every one of*
Existential quantifiers	*some, at least, exactly one, all but one*
Negation	*not every, no, not*
Adjectives	*green, happy, sad, good, bad, an abnormal,* and *an elegant*
Adverbs	*abnormally, elegantly*

Table 2: List of modifiers used to modify subject, verb, and object elements of sentences. They are applied to each of the premise and hypothesis sentences in Table 1.

subject, verb, object, subject + object, and verb + object.

3.3 Entailment Annotation Strategy

To annotate our dataset,[1] we created a set of annotation guidelines that follow Natural Logic and monotonicity calculus (MacCartney, 2009).

In general, to produce entailment relations we used a set theoretic approach to understand how the set of concepts that holds true in the premise overlaps with the set described in the hypothesis. To implement this set theoretic approach consistently, we defined the quantitative interpretation for several more ambiguous modifiers such as *all but one*, *all*, *not every* as follows:

1. For the modifier *all*, we consider the size of the set of elements X to be greater than 0: $|X| > 0$. For example, in the case of the phrase *all children*, we consider the size of the set of children to be greater than 0.

2. For the *all but one* modifier, we consider the size of *all* as N and the size of *all but one* to be $N - 1$. Note that the size of *all but one* could thus theoretically be 0, when $N = 1$.

3. For *not every* we consider the size of the corresponding set X to be 0 or larger: $|X| \geq 0$ where X is any set defined over the sentence. *not every man* would make X as a set of all men but there exists zero or one or more men that would not be included in this set.

4. When we cannot determine the size of the intersection of the two sets of premise and hypothesis, we resolved the annotation to be a Neutral label among all 7 entailment relations.

5. When comparing quantifiers between modified premise, and hypothesis sentence pairs, we denote the sizes of sets mathematically for $P \cup H$, $P \cap H$, and the Universal set. For example, consider the premise: ***every turtle is following the fish*** and the hypothesis: ***every fish is following the turtle***. The set over the premise is P : $\forall X \in$ all turtles following one fish, and the set over hypothesis is H : $\forall X \in$

Entailment Relation	Set Theoretic Notation	Examples using WordNet Hierarchy
Equivalence	$X \equiv Y$	*couch* $\equiv$ *sofa*
Forward Entailment (FE)	$X \subset Y$	Hyponym: *crow* $\subset$ bird.
Reverse Entailment (RE)	$X \supseteq Y$	Hypernym: *Asian* $\supseteq$ *Thai*.
Negation (Neg)	$X \cap Y = \phi \wedge X \cup Y = \mathbb{U}$	Antonym: *able* $\neg$ *unable*
Alternation	$X \cap Y = \phi \wedge X \cup Y \neq \mathbb{U}$	Typically caused concepts with a shared hypernym: *cat* $\parallel$ *dog*. The corresponding hierarchy is : *carnivore* $\rightarrow$ *feline*, *canine*; *feline* $\rightarrow$ *cat*; and *canine* $\rightarrow$ *dog*.
Cover	$X \cap Y \neq \phi \wedge X \cup Y = \mathbb{U}$	*animal* $\smile$ *non-ape*
Independence	all other cases	*hungry* $\parallel$ *hippo*

Table 3: Entailment relations as defined in (MacCartney, 2009) with explanations using the WordNet hierarchy (Miller, 1995).

Premise	Hypothesis	SVO	Modifier Type	Label
an old man is sitting in a field	a man is sitting in a field	None	None	FE
every old man is sitting in a field	a man is sitting in a field	Subject	Universal	FE
an old man is sitting in a field	**every** man is sitting in a field	Subject	Universal	RE
an old man is **elegantly** sitting in a field	a man is **elegantly** sitting in a field	Verb	Adverb	FE
an old man is sitting in **every** field	a man is sitting in a field	Object	Universal	FE
an old man is sitting in a field	a man is sitting in **every** field	Object	Universal	Neutral

Table 4: Premise, hypothesis examples where one or both of the premise and hypothesis were modified. The text in bold indicates the change from the original text. The *SVO* column indicates the part of the sentence that was modified: subject, verb, or object (SVO). The Modifier type indicates which type of modifier was used to modify the parts of sentences. The label is the Entailment relation annotated by the annotators over modified data.

all fishes following the turtle. Thus, $P \cap H = \phi$. In this case, the label is Negation (see Table 3). Table 4 includes more examples with the corresponding entailment labels.

A total of 1,304 modified premise and hypothesis sentence pairs along with original sentence pairs were included in the final SICCK dataset. The data was annotated by 5 annotators which were distributed between two sub-groups of annotators, based on the complexity of the labels. In the first two rounds of annotations, we re-grouped to develop concrete guidelines for annotations, without defining too strict rules by leaving room for more natural "if-this-then-that" deductions. There were disagreements between annotations which were resolved by verifying the sizes of sets mathematically over $X \cup Y$, $X \cap Y$ to follow the entailment relations defined as in (MacCartney, 2009). While in the initial round the inter-annotator agreement was low ($k < 0.4$), the annotations were revised until each group of annotators converged.

Tables 5, 6, and 7 provide summary statistics about the SICCK dataset.

Modifier type	# of sentence pairs
With universal quantifiers	217
With existential quantifiers	303
With negation	167
With adjectives/adverbs	602

Table 5: Sentence counts in SICCK based on types of modifiers.

SVO modified	# of sentence pairs
Subject	560
Verb	220
Object	509

Table 6: Sentence counts in SICCK based on which syntactic structures are modified.

4 Evaluation

We conducted an evaluation of how NLI methods capture the explicit compositionality in our dataset using two configurations: a zero-shot setting, in which we used NLI systems trained externally, and a fine-tuned setting, in which the same models were fine-tuned using our dataset.

Entailment relations	# of sentence pairs
Forward Entailment	223
Reverse Entailment	27
Alternation	121
Negation	54
Negation|Alternation	260
Neutral	393
Equivalence	7
Cover	1
Cover|FE	1

Table 7: Label counts in SICCK. Note that Negation|Alternation indicates ambiguous labels where the two annotators did not converge.

4.1 Zero-shot Analysis of NLI Models

For this analysis, we evaluate three pretrained neural entailment models on our dataset. However, all these systems emit just the three "traditional" entailment labels (Forward Entailment, Contradiction, and Neutral) whereas our dataset contains the seven labels from NL. To align these label spaces, we performed the following transformations:

1. In case a system produces a Neutral label, we run the prediction in the opposite direction, i.e., from hypothesis to premise. If the top label in the reversed direction is Forward Entailment (FE), we label the pair as Reverse Entailment. Otherwise, we keep the Neutral label. This heuristic allows these systems to produce four labels instead of three.

2. We convert our seven labels to four labels through the following heuristics: (a) Equivalence was removed since we had only one sentence pair labeled as Equivalence in our dataset; (b) Alternation is merged with Negation; (c) Cover and Independence become Neutral; and (d) the 7 examples that were annotated as Cover|FE were removed.

We conducted zero shot evaluation using three NLI models: the cross-encoder model of Reimers and Gurevych (2019) (`nli-deberta-v3-base` in our tables), the adversarial NLI model of Nie et al. (2020) (`ynie/roberta-large-...`), and ELMo-based Decomposable Attention model (Parikh et al., 2016) (`pair-classification-...`). We draw the following observations from this experiment:

- Table 8 indicates that the ELMO-based NLI model performs considerably worse than the other two transformer-based models. This is a testament to how far our field has progressed in just a few years. However, no model approaches 70 F1 points, which indicates that none of these models truly understand the task well.

- The NLI models do better over adjectives and adverbs, but they struggle to understand statements modified with universal and existential quantifiers, and negation. Tables 8–14 indicates that the transformer-based NLI models perform at over 70 F1 points on adjectives/adverbs, at over 65 F1 for universal quantifiers, at approximately 60 F1 for existential quantifiers, and at only 30–35 F1 for negation. This is a surprising finding considering how much attention negation has received in the NLP literature (Pang et al., 2002) (Hossain et al., 2022) (Hossain et al., 2020).

- Lastly, Tables 15–17 indicate that NLI models process objects best, followed by subjects, and, lastly, verbs. This is not surprising considering the increased semantic ambiguity of verbs.

4.2 Analysis of Fine-tuned NLI models

To understand if NLI methods are capable of learning this compositional information, we fine-tuned the two NLI models that performed better over the SICCK dataset. To maximize the data available, we implemented a 5-fold cross-validation evaluation over the entire SICCK dataset and experimented with multiple hyperparameters. In particular, we used 4 or 8 epochs, and batch sizes of 8, 16, or 32 data points.

The results of these experiments are summarized in Table 9. We draw the following observation from this experiment:

- The difference in F1 scores between the fine-tuned systems and the corresponding zero-shot setting ranges from -0.05 to 0.06. This indicates that these systems do not acquire substantial new knowledge despite the fact that they've been exposed to approximately 1,300 sentences with compositional information. This suggests that understanding compositionality is harder than expected.

- Similar to the zero-shot setting, NLI models did better over adjectives, and adverbs and

NLI system	F1	Precision	Recall	Accuracy
nli-deberta-v3-base	**0.5254**	**0.5601**	**0.5860**	**0.6579**
ynie/roberta-large-snli-mnli-fever-anli-R1-R2-R3-nli	0.5200	0.5156	0.5533	0.6334
pair-classification-decomposable-attention-elmo	0.0829	0.0497	0.2500	0.1986

Table 8: Overall scores for the three pretrained NLI modes under zero-shot setting, based on compressed 4-entailment relations: Forward Entailment, Reverse Entailment, Contradiction, and Neutral.

NLI model with epochs, batch size	F1	Precision	Recall	Accuracy
nli-deberta-v3-base-4-8	0.5392	0.5310	0.5546	0.5686
roberta-large-snli-mnli-fever-anli-R1-R2-R3-nli-4-8	0.4257	0.4708	0.4456	0.4460
nli-deberta-v3-base-4-16	0.5153	0.5072	0.5404	0.5556
roberta-large-snli-mnli-fever-anli-R1-R2-R3-nli-4-16	0.5177	0.5472	0.5239	0.5203
nli-deberta-v3-base-4-32	0.4871	0.4756	0.5109	0.5280
roberta-large-snli-mnli-fever-anli-R1-R2-R3-nli-4-32	0.4829	0.4941	0.4884	0.4904
nli-deberta-v3-base-8-8	**0.5876**	**0.5799**	**0.6024**	**0.6045**
roberta-large-snli-mnli-fever-anli-R1-R2-R3-nli-8-8	0.4433	0.4596	0.4385	0.4559
nli-deberta-v3-base-8-16	0.5464	0.5378	0.5636	0.5801
roberta-large-snli-mnli-fever-anli-R1-R2-R3-nli-8-16	0.4724	0.4723	0.4772	0.4828
nli-deberta-v3-base-8-32	0.5489	0.5408	0.5713	0.5785
roberta-large-snli-mnli-fever-anli-R1-R2-R3-nli-8-32	0.4281	0.4558	0.4319	0.4345

Table 9: Overall scores for two *fine-tuned* NLI models on SICCK dataset based on the compressed 4-entailment relations: Forward Entailment, Reverse Entailment, Contradiction, and Neutral. We report the results for 4 and 8 epochs and batch sizes of 8, 16, and 32 data points.

relatively better over universal quantifiers in comparison to that of the negation and existential quantifiers. We also observed that models seem to be confused when the annotated label was Neutral but the modifier types were negations.

- NLI models perform somewhat better over subject and object-modified examples than on examples with modified verbs. This indicates that the semantic ambiguity of verbs is likely to impact NLI models.

5 Error Analysis

We analyze the incorrect predictions of the NLI models over SICCK dataset in this section. We observed that NLI models performed better over adjectives and adverbs, and relatively well over universal quantifiers in comparison to sentences modified with negation and existential quantifiers under both fine-tuned as well as zero-shot settings. We also observed that models seem to be confused when the adjusted label was Neutral and the modifier types were negations.

SVO	# count	Neutral
subject	86	65
verb	41	35
object	40	22

Table 10: For all our SICCK dataset's 167 examples with negation modifiers, this table includes counts of all the modified subject, verb, and object parts of sentences respectively for each of the 4-Entailment adjusted labels from SICCK annotations. The last column indicates how many of these data points have the Neutral label.

5.1 Neutral Labels with Negation Modifiers

Negation understanding in natural language has been a challenging problem (Pang et al., 2002; Hossain et al., 2022). (Hossain et al., 2022) discussed that Negation is underrepresented in natural language corpora. Further, (Hossain et al., 2020) show that even though the transformers were fine-tuned with modified premises with negation (i.e., verb modifiers with negation), the transformers struggle with inference over negated sentence pairs.

In our SICCK dataset, there are 167 examples with negation modifiers. Table 10 shows some statistics relevant to this. Of these 167 examples with Negation modifiers, there are 122 Neutral ex-

amples. We observed that the NLI models correctly predicted ground truth for approximately 37% of these examples. For all the incorrectly predicted labels for negation-modified examples with Neutral labels, the models seemed to be confused for various compositional cases, i.e. subject or verb or object-modified examples almost equivalently. Modifiers such as *no*, *not every*, *not* with Neutral and Contradiction labels seem to contribute to the confusion. SICCK examples also include the format of alternating modifiers between premises, hypothesis, or both i.e. $(P_i', H_i), (P_i, H_i'), (P_i', H_i')$ Section 3 which further seems to confuse the NLI models. This is surprising since we have 593 Neutral examples in our SICCK dataset, albeit with fewer negation examples. As emphasized by the analysis from (Hossain et al., 2022) and (Hossain et al., 2020), detecting negation in natural language continues to be an unresolved problem.

5.2 Verb-modified Examples

For verb modifiers, we used modifiers *abnormally*, *elegantly*, *always*, *never*. Our SICCK dataset has a total of 220 verb-modified examples of which, we have 89 universal modifiers, 90 adverbs/adjectives, and 41 negation. Among the 31 verb-modified examples with negation with Neutral label, NLI models incorrectly alternate between Contradiction and FE for 99% of the examples. Of the 49 examples with universal modifiers over verbs with Neutral labels, approximately 69.4% were incorrectly predicted. This further emphasizes that negation (especially when occurring in Neutral examples) remains a challenge.

6 Conclusion

This paper introduced a new, synthetic dataset that facilitates analyses of how NLI models capture compositionality. The dataset contains 1,304 sentence pairs that were created by modifying 15 examples from the SICK dataset (Marelli et al., 2014) with a variety of modifiers that correspond to universal quantifiers, existential quantifiers, negation, and other concept modifiers in Natural Logic (NL) (MacCartney, 2009). We used these phrases to modify the subject, verb, and object parts of the premise and hypothesis. Lastly, we annotated these modified texts with the corresponding entailment labels following NL rules.

We conducted a preliminary analysis of how well the change in the structural and semantic composi-

tion is captured and detected by neural NLI models, in both zero-shot and fine-tuned scenarios. We found that the performance of NLI models is poor in both settings, especially for modified sentences with negation and existential quantifiers, and when verbs are modified.

Limitations

While this work explores the impact of the typical compositional modifiers on entailment relations, we did not consider other fine-grained information that further captures upward or downward monotonicity from the monotonicity calculus of the premise/hypothesis sentence pairs. Further, the dataset that we generated is relatively small, at approximately 1,300 sentences. We also did not evaluate the dataset over T5, BART, GPT-x, and other state-of-the-art LLMs, which may provide more insights. We also did not conduct any evaluation for explanations and interpretation of the evaluated NLI models, which could be future work. Lastly, we did not include a comparison with existing datasets that were created specifically for negation modifiers and universal & existential quantifiers. We see all these issues as exciting avenues for future work.

References

Lasha Abzianidze and Johan Bos. 2017. Towards universal semantic tagging. *arXiv preprint arXiv:1709.10381*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015a. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015b. The snli corpus.

Samuel R. Bowman, Christopher Potts, and Christopher D. Manning. 2015c. Recursive neural networks can learn logical semantics.

Robin Cooper, R Crouch, Jan van Eijck, Chris Fox, Josef van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. Using the framework. the fracas consortium. Technical report, Technical report, FraCaS deliverable D-16.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J Gershman, and Noah D Goodman. 2018.

Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Maayan Geffet and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 107–114.

Izumi Haruta, Koji Mineshima, and Daisuke Bekki. 2020. Logical inferences with comparatives and generalized quantifiers.

Md Mosharaf Hossain, Dhivya Chinnappa, and Eduardo Blanco. 2022. An analysis of negation in natural language understanding corpora. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 716–723, Dublin, Ireland. Association for Computational Linguistics.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

Hai Hu, Qi Chen, Kyle Richardson, Atreyee Mukherjee, Lawrence S Moss, and Sandra Kuebler. 2019. Monalog: a lightweight system for natural language inference based on monotonicity. *arXiv preprint arXiv:1910.08772*.

Rohan Jha, Charles Lovering, and Ellie Pavlick. 2020. Does data augmentation improve generalization in nlp? *arXiv preprint arXiv:2004.15012*.

Pratik Joshi, Somak Aditya, Aalok Sathe, and Monojit Choudhury. 2020. Taxinli: Taking a ride up the nlu hill. *arXiv preprint arXiv:2009.14505*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Bill MacCartney. 2009. *Natural language inference*. Stanford University.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In

Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.

Ankur P. Parikh, Oscar T"ackstr"om, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. *ArXiv*, abs/1606.01933.

Abhilasha Ravichander, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. 2019. Equate: A benchmark evaluation framework for quantitative reasoning in natural language inference.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Kyle Richardson, Hai Hu, Lawrence Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8713–8721.

Julia Rozanova, Deborah Ferreira, Marco Valentino, Mokanrarangan Thayaparan, and Andre Freitas. 2021. Decomposing natural logic inferences in neural nli. *arXiv preprint arXiv:2112.08289*.

Mark Steedman and Jason Baldridge. 2011. Combinatory categorial grammar. *Non-Transformational Syntax: Formal and Explicit Models of Grammar. Wiley-Blackwell*, pages 181–224.

Robert Vacareanu, George Caique Gouveia Barbosa, Marco A. Valenzuela-Escárcega, and Mihai Surdeanu. 2020. Parsing as tagging. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5225–5231, Marseille, France. European Language Resources Association.

Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. 2021a. Info{bert}: Improving robustness of language models from an

information theoretic perspective. In *International Conference on Learning Representations*.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021b. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019a. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019b. Help: A dataset for identifying shortcomings of neural models in monotonicity reasoning. *arXiv preprint arXiv:1904.12166*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding.

A Appendix

We provide the modifier-based evaluation results for zero-shot setting over NLI models and finetuned NLI models.

NLI system	F1	Precision	Recall	Accuracy
nli-deberta-v3-base	0.5333	0.5474	**0.5877**	0.6636
ynie/roberta-large-snli-mnli-fever-anli-R1-R2-R3-nli	**0.5597**	**0.5636**	0.5854	**0.6774**
pair-classification-decomposable-attention-elmo	0.0694	0.0403	0.2500	0.1613

Table 11: **Universal quantifiers**: scores based on compressed 4-entailment relations for zero-shot NLI evaluation.

NLI system	F1	Precision	Recall	Accuracy
nli-deberta-v3-base	**0.5101**	**0.5453**	**0.5608**	**0.6106**
ynie/roberta-large-snli-mnli-fever-anli-R1-R2-R3-nli	0.4932	0.4888	0.5348	0.5941
pair-classification-decomposable-attention-elmo	0.1044	0.0660	0.2500	0.2640

Table 12: **Existential quantifiers**: scores based on compressed 4-entailment relations for zero-shot NLI evaluation.

NLI system	F1	Precision	Recall	Accuracy
nli-deberta-v3-base	0.2678	**0.3556**	0.4637	**0.3054**
ynie/roberta-large-snli-mnli-fever-anli-R1-R2-R3-nli	**0.2996**	0.3470	**0.4690**	0.3533
pair-classification-decomposable-attention-elmo	0.0412	0.0225	0.2500	0.0898

Table 13: **Negation**: scores based on compressed 4-entailment relations for zero-shot NLI evaluation.

NLI system	F1	Precision	Recall	Accuracy
nli-deberta-v3-base	**0.5768**	**0.6088**	**0.6154**	**0.7741**
ynie/roberta-large-snli-mnli-fever-anli-R1-R2-R3-nli	0.5543	0.5618	0.5523	0.7126
pair-classification-decomposable-attention-elmo	0.1139	0.0687	0.3333	0.2060

Table 14: **Adjectives/Adverbs**: scores based on compressed 4-entailment relations for zero-shot NLI evaluation.

NLI system	F1	Precision	Recall	Accuracy
nli-deberta-v3-base	**0.5070**	**0.5469**	**0.5732**	**0.6429**
ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli	0.4862	0.4812	0.5168	0.6054
pair-classification-decomposable-attention-elmo	0.0796	0.0473	0.2500	0.1893

Table 15: **Modified subject**: scores based on compressed 4-entailment relations for zero-shot NLI evaluation.

NLI system	F1	Precision	Recall	Accuracy
nli-deberta-v3-base	0.4724	**0.5141**	**0.5875**	0.5727
ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli	**0.4932**	0.5082	0.5764	**0.5955**
pair-classification-decomposable-attention-elmo	0.0669	0.0386	0.2500	0.1545

Table 16: **Modified verb**: scores based on compressed 4-entailment relations for zero-shot NLI evaluation.

NLI system	F1	Precision	Recall	Accuracy
nli-deberta-v3-base	0.5683	**0.6166**	**0.6030**	**0.7073**
ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli	**0.5687**	0.5638	0.5908	0.6778
pair-classification-decomposable-attention-elmo	0.0915	0.0560	0.2500	0.2240

Table 17: **Modified object**: scores based on compressed 4-entailment relations for zero-shot NLI evaluation.

NLI system	F1	Precision	Recall	Accuracy
nli-deberta-v3-base	**0.5393**	**0.5319**	**0.5542**	**0.5622**
ynie/roberta-large-snli-mnli-fever-anli-R1-R2-R3-nli	0.3960	0.4241	0.4046	0.3917

Table 18: **Universal quantifiers**: fine-tuned NLI models' evaluation scores based on 4-entailment relations. We report results for 8 epochs and a batch size of 32 data points.

NLI system	F1	Precision	Recall	Accuracy
nli-deberta-v3-base	**0.5470**	**0.5373**	**0.5729**	**0.5611**
ynie/roberta-large-snli-mnli-fever-anli-R1-R2-R3-nli	0.4649	0.5101	0.4633	0.4653

Table 19: **Existential quantifiers**: fine-tuned NLI models' evaluation scores based on 4-entailment relations. We report results for 8 epochs and a batch size of 32 data points.

NLI system	F1	Precision	Recall	Accuracy
nli-deberta-v3-base	0.3476	0.3634	0.4064	0.3952
ynie/roberta-large-snli-mnli-fever-anli-R1-R2-R3-nli	**0.3979**	**0.4420**	**0.4536**	**0.4671**

Table 20: **Negation**: fine-tuned NLI models' evaluation scores based on 4-entailment relations. We report results for 8 epochs and a batch size of 32 data points.

NLI system	F1	Precision	Recall	Accuracy
nli-deberta-v3-base	**0.5862**	**0.5803**	**0.5998**	**0.6412**
ynie/roberta-large-snli-mnli-fever-anli-R1-R2-R3-nli	0.4191	0.4350	0.4201	0.4252

Table 21: **Adjectives/Adverbs**: fine-tuned NLI models' evaluation scores based on 4-entailment relations. We report results for 8 epochs and a batch size of 32 data points.

NLI system	F1	Precision	Recall	Accuracy
nli-deberta-v3-base	**0.5920**	**0.5848**	**0.6154**	**0.6107**
ynie/roberta-large-snli-mnli-fever-anli-R1-R2-R3-nli	0.4681	0.4872	0.4778	0.4768

Table 22: **Modified subject**: fine-tuned NLI models' evaluation scores based on 4-entailment relations. We report results for 8 epochs and a batch size of 32 data points.

NLI system	F1	Precision	Recall	Accuracy
nli-deberta-v3-base	**0.5203**	**0.5114**	**0.5454**	**0.5545**
ynie/roberta-large-snli-mnli-fever-anli-R1-R2-R3-nli	0.3750	0.4102	0.3864	0.3955

Table 23: **Modified verb**: fine-tuned NLI models' evaluation scores based on 4-entailment relations. We report results for 8 epochs and a batch size of 32 data points.

NLI system	F1	Precision	Recall	Accuracy
nli-deberta-v3-base	**0.5501**	**0.4945**	**0.5342**	**0.5058**
ynie/roberta-large-snli-mnli-fever-anli-R1-R2-R3-nli	0.4047	0.4410	0.4011	0.4053

Table 24: **Modified object**: fine-tuned NLI models' evaluation scores based on 4-entailment relations. We report results for 8 epochs and a batch size of 32 data points.