Synthetic Dataset for Evaluating Complex Compositional Knowledge for Natural Language Inference

Sushma Anand Akoju Robert Vacareanu Haris Riaz
Eduardo Blanco Mihai Surdeanu

Computational Language Understanding Lab at the University of Arizona

SICCK Dataset Examples

The SICCK dataset contains premise/hypothesis sentences changed with logic modifiers

Premise	Hypothesis	SVO	Modified Premise/Hypothesis/both	Modifier	Modifier Type	Label
an old man is sitting in a field	a man is sitting in a field	None	None	None	None	FE
an old man is sitting in a field	**every** man is sitting in a field	SUBJ	Hypothesis	every	Universal	RE
an old man is **never** sitting in a field	a man is sitting in a field	VERB	Premise	never	Universal	Neutral
an old man is sitting in a field	a man is sitting in **every** field	OBJ	Hypothesis	every	Universal	Neutral
some old man is sitting in a field	a man is sitting in a field	SUBJ	Premise	some	Existential	FE
an old man is sitting in **some** field	a man is sitting in a field	OBJ	Premise	some	Existential	FE
not every old man is sitting in a field	a man is sitting in a field	SUBJ	Premise	not every	Negation	Neutral
an old man is sitting in a field	a man is **not** sitting in a field	VERB	Hypothesis	not	Negation	Contradiction
an old man is sitting in **no** field	a man is sitting in **no** field	OBJ	Premise	no	Negation	FE
a **bad** old man is sitting in a field	a **bad** man is sitting in a field	SUBJ	Both	bad	Adverb/**Adjective**	FE
an old man is **elegantly** sitting in a field	a man is **elegantly** sitting in a field	VERB	Both	elegantly	**Adverb**/Adjective	FE
an old man is sitting in a field	a man is sitting in **an abnormal** field	OBJ	Hypothesis	an abnormal	Adverb/**Adjective**	Neutral

Table 1: Premise, hypothesis examples where one or both of the premise and hypothesis, SVO, Modifier type were modified. SVO indicates the part of the sentence that was modified i.e subject, verb, or object. The Modifier type indicates one of the 4 types of modifiers used to modify the parts of sentences. Labels are 4-entailment relations: Forward Entailment (FE), Reverse Entailment (RE), Contradiction, and Neutral.

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 . 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). 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.

SICCK Modifiers

Modifiers Used to Construct SICCK

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

Table 2: List of modifiers used to modify SUBJ-VERB-OBJ elements of sentences.

Approach

Key Components

- **Sentence Modification:** we produce 1304 examples from 15 SICK premise, and 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`
- **Annotation *guidelines* :** annotation guidelines based on monotonicity calculus and natural logic for annotating the modified premise and hypothesis sentences.
- **Analysis of zero-shot and fine-tuned NLI models** indicates that these structural and compositional changes are not captured well by these models.

F1- Scores over Modifier Types

Finetuned NLI Models Continue to Perform Poorly, Especially on Negation Modifiers

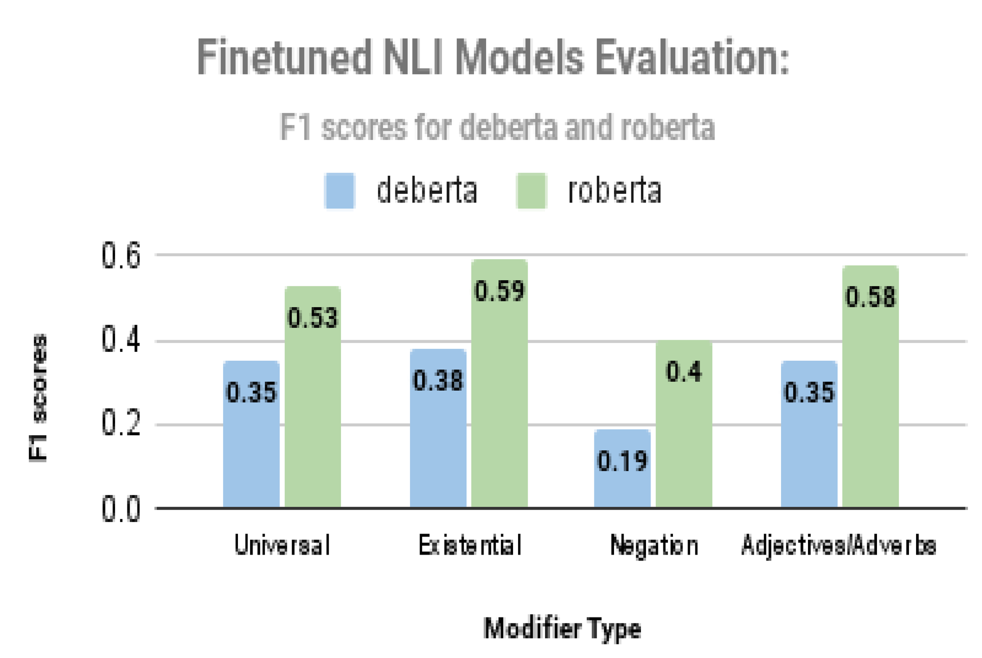

Figure 1: Finetuned NLI Models: F1- scores over Modifier type

Zero-shot Evaluation

NLI Models Perform Poorly in a Zero-shot Setting

NLI system	F1
deberta	**0.5254**
roberta-large	0.5200
elmo	0.0829

Table 3: Overall scores for the three pretrained NLI modes under zero-shot setting.

Fine-tuned Evaluation

NLI Models Perform only Marginally Better when Fine-tuned

NLI model with epochs, batch size	F1
roberta-large-4-8	(0.52±0.02)
deberta-4-8	(0.33±0.02)
roberta-large-4-16	(0.59±0.04)
deberta-4-16	(0.34±0.01)
roberta-large-4-32	**(0.62±0.04)**
deberta-4-32	(0.37±0.01)
roberta-large-8-8	(0.49±0.06)
deberta-8-8	(0.33±0.02)
roberta-large-8-16	(0.53±0.04)
deberta-8-16	(0.33±0.02)
roberta-large-8-32	(0.57±0.01)
deberta-8-32	(0.34±0.01)

Table 4: Overall scores for two *fine-tuned* NLI models on SICCK dataset.