**Hypothesis only baselines in Natural Language Inference**

**Negative**

The largest relative gains are on human- elicited models where the hypothesis-only model more than doubles the majority baseline

If p(l|w) is highly skewed across labels, there exists the potential for a predictive bias

All words are considered highly-correlated with a specific class label, and thus the entire data set would be treated as trivially answerable

As a common strategy, crowd-source workers often do not generate contradictory hypotheses that require fine-grained semantic reasoning, as a majority of such activities can be easily negated by removing an agent’s agency

Universal negation constitutes four of the remaining seven terms in this list, and may also be used to similar effect

A number of its property-driven hypotheses, such as *X was sentient in [the event]*, can be accu- rately guessed based on lexical semantics (back- ground knowledge learned from training) of the argument

**Positive**

The drop between SNLI and Multi-NLI suggests that by including multiple genres, an NLI dataset may contain less biases. However, adding additional genres might not be enough to mitigate biases as the hypothesis-only model still drastically outperforms the majority-baseline

Out of 50 sentences that the model correctly labeled as ENTAILED, 88% of them were grammatical. On the otherhand, of the 50 hypotheses incorrectly labeled as ENTAILED, only 38% of them were grammatical. Similarly, when the model correctly labeled 50 NOT-ENTAILED hypotheses, only 20% were grammatical, and 68% when labeled incorrectly.

**Annotation Artifacts in Natural Language Inference Data**

**Negative**

Generic words such as *animal*, *instrument*, and *outdoors*

Replace exact numbers with approximates (*some*, *at least*, *various*)

Remove explicit gender (*human* and *person*)

Modifiers (*tall*, *sad*, *popular*) and superlatives (*first*, *favorite*, *most*)

Cause and purpose clauses, which increase the prevalence of discourse markers such as *because*

*Negation words, Sleeping* contradicts any activity, and *naked* (further down the list) contradicts any description of clothing.

SNLI, neutral hypotheses tend to be long, while entailed ones are generally shorter

8.8% of entailed hypotheses in SNLI are fully contained within their premise, while only 0.2% of neutrals and contradictions exhibit the same property

**Positive**

Length is also a discriminatory feature in MultiNLI, but is less significant, possibly due to the introduction of diverse genres

**Probing Neural Network Comprehension of Natural Language Arguments**

**Negative**

Exploits the presence of cue words in the warrant, especially “not”

On small datasets, BERT sometimes fails to train, yielding degenerate results

Productivity of a cue measures the benefit of exploiting it, while coverage measures the strength of the signal it provides

**Positive**

**Formulae**

Applicability of cue: number of data points where it occurs with one label but not the other

A picture containing object

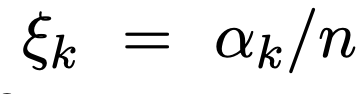
Description automatically generated

Productivity of cue: proportion of applicable data points for which it predicts the correct answer

A close up of a logo

Description automatically generated

Coverage of cue: proportion of applicable cases over the total number of data points



**Investigating Biases in Textual Entailment Datasets**

**Negative**

As one of the reasons for the NLI task was for the learning of sentence representations, we also trained an LSTM sentence-embedding encoder. The idea was to compare the performance between a model that uses a fixed-length sentence embed- ding and one that tries to model interactions be- tween hidden states of an RNN (ESIM and DIIN fit into this category). Because sentence embed- ding models do not force the ‘interaction’ between the two inputs, we believe that the sentence em- bedding models may be more prone to learning these superficial correlations

For SNLI, we find that the informative bigrams make up the long-tail of the bigram distributions, but many of them are predictive of the labels. MultiNLI also has many low frequency bi- grams that are preferentially predictive of contra- diction

**Positive**

During testing, we shuffle the premises so that they do not correspond to the right hypotheses. The sentence-embedding models that we trained achieved 70% accuracy when trained on the full dataset while under the shuffled premise test, they achieved an accuracy of 50%. In comparison, the ESIM model achieved a 40.5% accuracy in this setting. This suggests that the model still uses some of the correlations found in the hypothesis

**How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks**

**Negative**

nullifying the information present in either questions or passages: bAbi: task performance about the same

make use of just last sentence instead of all 20 sentences in the passage, our sentence memory based KV-MemNet achieve comparable or better performance *w.r.t* the *full* model on most subtasks: CBT

**Positive**

anonymization of entities which pre- vents models from building entity-specific information : CNN

**What Makes Reading Comprehension Questions Easier?**

**Negative**

SQAUD: easily attend to the answer- contained sentence (s1) by watching word overlaps

questions whose correct answer appears or does not appear in the context sentence that is most similar to the question

Entity recog heuristic:

Because the questions in QAngaroo are not complete sentences, but rather knowledge-base entries that have a blank k=1 has only 1.8 difference

1. the existence of a single candidate answer that is restricted by ex- pressions such as “wh-” and “how many”
2. lexical patterns that appear around the correct answer.

Attention heuristic:

1. how many questions have their correct answers in the most similar sentence
2. whether a performance gap exists for such questions

**Positive**

extended the notion of “sentence” in our annotation and considered a subordinate clause as a sentence. This modification was intended to deal with the internal complexity of a sentence with multiple clauses, which can also render a question difficult

ambiguous questions: with multiple correct spans - important feature insofar because it can lead to unstable scoring in EM/F1

easier questions are those (i) where the reader needs to generate only one hypothesis, and (ii) where the premises di- rectly describe the correct hypothesis

integration of premises should be complemented by external knowledge to provide sufficient information to verify the cor- rect hypothesis

**oLMpics - On what Language Model Pre-training Captures**

**Negative**

LM objective focuses on word co-occurrence

For example, ROBERTA-L can almost perfectly compare people’s ages, when the numeric values are in the expected range (15-105), but miserably fails if the values are outside this range

**Positive**

LM will struggle with tasks that are considered to involve symbolic reasoning such as determining whether a *conjunction* of properties is held by an object, and *comparing* the sizes of different ob- jects

one task that focuses on negation uses a the template *“It was [MASK] fat, it was really slim”*, with candidate answers *“not”*, *“very”*. Such sentences may derail a LM due to the unnatural language rather than the ability to interpret negation

replace the word *“and”* with the word *“blah”*, resulting in examples such as *“What is located at hand blah used for writing?* : good perf means ‘and’ is not used in reasoning

Adversarial Filters of Dataset Biases

**Negative**

biases or *artifacts* are often introduced during data collection or human annotation

high predictability scores are undesirable as their feature representation can be exploited to confidently correctly predict such instances

AFLITE substantially outperforms challenging examples on the HANS bench- mark, which targets models purely relying on lexical and syntactic cues.

**Positive**

**Formulae**

subsets T =(XT,YT)of S

*predictability score* p(i) for i: on average, how reliably can label yi be predicted using features Φ(xi) when a model from M is trained on a randomly chosen training set S \ T not containing i

**Stress Test Evaluation for Natural Language Inference**

**Negative**

See error section

**Positive**

premise-hypothesis pairs in inference usually consist of a single sentence, but concatenative adversaries break this assumption

See error section

**A MUTUAL INFORMATION MAXIMIZATION PERSPEC- TIVE OF LANGUAGE REPRESENTATION LEARNING**

**Negative**

Image representation learning methods often incorporate a regularization term in its objective function to encourage learned representations to look like a prior distribution. This is useful for incorporating prior knowledge into a representation learning model.

regularization

**Positive**

**What Knowledge is Needed to Solve the RTE5 Textual Entailment Challenge?**

**Negative**

**T:** ...The cap for a "roaming" text will fall...from about 29 cents today.... **H:** A roaming text cost 46 euro cents. : can’t do with lexical overlap

"The ban..." in H presupposes there is a ban, rather than asks if there was one

**Positive**

we know that one way MPs can cut costs is to impose a cap. To use this, we'd also need to relate "texting" to "a text".

metonymy: "The cap for a text" means "The cap for *sending* a text" (you pay to send a text, not to obtain ownership of it).

*elaboration* occurs, we have to align the elements in it with *introduction*, even if there is no definite reference: use coreference and pragmatics

jury is part of judging

"A...man..admitted killing...” → “A man killed..."

a Russian cosmonaut, an American astronaut and U.S. billionaire tourist Charles Simonyi...

Implicit noun "more than 30" → "more than 30 people"

summing the numbers "10" + "more than 30" to conclude "more than 40"

"in the capital" → in Baghdad → in Iraq

Theme: Morphosemantics : We also need to know that if you sell something, you are the seller (morphosemantics). {seller,vendor} are synonyms in WordNet.

The U.S...is working to dissuade North Korea...[from] launch[ing] a satellite", suggesting the U.S. fears the launch.

World Moto GP champion is unnatural and incorrectly bracketed as champion OF (all 3)

Common scripts: plot lines in the stories

**Doubt**

Theme: Grain size/granularity reasoning : We also need: IF the Chinese government does X THEN China does X

**GENERATING NATURAL ADVERSARIAL EXAMPLES**

**Negative**

replace tokens by random words of the same POS tag with probability proportional to embedding similarity.

**Positive**

Negative context words increase: “There are children present” is entailed by the sentence “Children smiling and waving at camera”, while the sentence “The kids are frowning” contradicts it. We use our approach to generate adversaries by perturbing the hypothesis to deceive classifiers, keeping the premise unchanged.

Table 12: “Adversaries” that find dropped verbs in English-To-German translation.

Generated Translation (German)

s: A man looks back while laughing and walking.

s′ : A man is laughing walking down the ground

Ein Mann schaut beim Lachen und Gehen zurck.

Ein Mann lacht auf dem Boden.*(A man laughs on the floor.)*

s: She is cooking food while wearing a dress. s′ : She is cooking dressed for a wedding.

Sie kocht Essen, whrend sie ein Kleid trgt.

Sie kocht fr eine Hochzeit. *(She cooks for a wedding.)*

**Explain Yourself! Leveraging Language Models for Commonsense Reasoning**

**Negative**

Lexical overlay q and a

**Positive**

less successful on tasks that require clear understanding of how pronouns resolve

generated explanations exhibited little grammatical or syntactical errors and often contained apparently relevant information.

**What Does My QA Model Know? Devising Controlled Probes using Expert Knowledge**

**Negative**

Word net qa isa based triple inputs

**Positive**

general knowledge about word definitions and general taxonomic reasoning which are fundamental to more complex forms of reasoning and are widespread in benchmark datasets : hypernymy, hyponymy, and synonymy detection, and word sense disambiguation

question *perturbations* (i.e., systematic adjustments to how the questions are constructed)

*the quickness with which they are able to learn the inoculated task provides evidence of prior competence*, which is precisely what we aim to probe. To measure past perfor- mance, we define a model’s inoculation cost as the difference in the performance of this model on its original task before and after inoculation.

**Co-opNet: Cooperative Generator–Discriminator Networks for Abstractive Summarization with Narrative Flow**

**Negative**

n-gram overlap without distinguishing between salient n-grams and non-contentful tokens.

**Positive**

scientific papers incorporating a wide range of domain knowledge and subjects to rigorously test the generalization of our models across different fields of scientific endeavour.

* Length of summaries → Are the summaries long enough to clearly show narrative flow properties?
* Abstractiveness of gold summaries → Do the summaries exhibit particular sentence-level flow, or are the summary sentences extracted highlights from the context?

**Exploring Numeracy in Word Embeddings**

**Negative**

contexts that separates numbers of vastly varying magnitudes, we sample 1 million sentences con- taining numbers from English Wikipedia and Gi- gaWord and compute pointwise mutual informa- tion (PMI)

**Positive**

**Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data**

**Negative**

**Positive**

**﻿Semi-Autoregressive Training Improves Mask-Predict Decodin﻿g**

**Negative**

**Positive**

We first create training examples by starting with the gold target sequence and masking a subset of its tokens, just like the original training process. We then use the current model to predict the sequence from the partially-observed input, and mask a different subset of tokens to create the training example’s input. The model is then trained to predict the gold target sequence based on this partially-observed prediction-based input, as well as the source se- quence (see Figure 1), allowing it to better correct mistakes made during the early iterations of the mask-predict decoding loop.

**Fine-Tuning a Transformer-Based Language Model to Avoid Generating Non-Normative Text**

**Negative**

**Positive**