Hypothesis Only Baselines in Natural Language Inference

Adam Poliak[1] **Jason Naradowsky**[1] **Aparajita Haldar**[1,2]
Rachel Rudinger[1] **Benjamin Van Durme**[1]
[1]Johns Hopkins University [2]BITS Pilani, Goa Campus, India
{azpoliak,vandurme}@cs.jhu.edu {narad,ahaldar1,rudinger}@jhu.edu

Abstract

We propose a *hypothesis only* baseline for diagnosing Natural Language Inference (NLI). Especially when an NLI dataset assumes inference is occurring based purely on the relationship between a context and a hypothesis, it follows that assessing entailment relations while ignoring the provided context is a degenerate solution. Yet, through experiments on ten distinct NLI datasets, we find that this approach, which we refer to as a hypothesis-only model, is able to significantly outperform a majority-class baseline across a number of NLI datasets. Our analysis suggests that statistical irregularities may allow a model to perform NLI in some datasets beyond what should be achievable without access to the context.

1 Introduction

Though datasets for the task of Natural Language Inference (NLI) may vary in just about every aspect (size, construction, genre, label classes), they generally share a common structure: each instance consists of two fragments of natural language text (a *context*, also known as a *premise*, and a *hypothesis*), and a label indicating the entailment relation between the two fragments (e.g., ENTAILMENT, NEUTRAL, CONTRADICTION). Computationally, the task of NLI is to predict an entailment relation label (output) given a premise-hypothesis pair (input), i.e., to determine whether the truth of the hypothesis follows from the truth of the premise (Dagan et al., 2006, 2013).

When these NLI datasets are constructed to facilitate the training and evaluation of natural language understanding (NLU) systems (Nangia et al., 2017), it is tempting to claim that systems achieving high accuracy on such datasets have successfully "understood" natural language or at least a logical relationship between a premise and hypothesis. While this paper does not attempt to prescribe the sufficient conditions of such a claim, we argue for an obvious *necessary*, or at least *desired* condition: *that interesting natural language inference should depend on both premise and hypothesis.* In other words, a baseline system with access only to hypotheses (Figure 1b) can be said to perform NLI only in the sense that it is understanding language based on prior background knowledge. If this background knowledge is about the world, this may be justifiable as an aspect of natural language understanding, if not in keeping with the spirit of NLI. But if the "background knowledge" consists of learned statistical irregularities in the data, this may not be ideal. Here we explore the question: do NLI datasets contain statistical irregularities that allow hypothesis-only models to outperform the datasets specific prior?

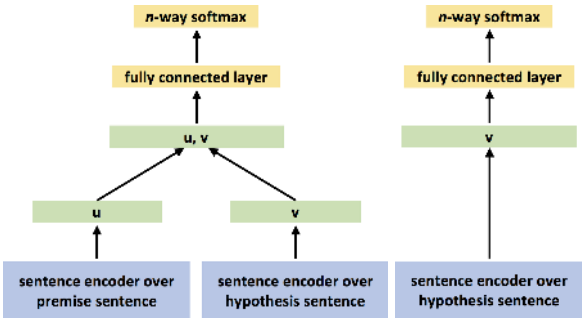


Figure 1: (1a) shows a typical NLI model that encodes the premise and hypothesis sentences into a vector space to classify the sentence pair. (1b) shows our hypothesis-only baseline method that ignores the premise and only encodes the hypothesis sentence.

We present the results of a hypothesis-only baseline across ten NLI-style datasets and advocate for its inclusion in future dataset reports. We find that this baseline can perform above the majority-class prior across most of the ten examined datasets. We examine whether: (1) hypotheses contain statistical irregularities within each

*Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 180–191
New Orleans, June 5-6, 2018.

entailment class that are "giveaways" to a well-trained hypothesis-only model, (2) the way in which an NLI dataset is constructed is related to how prone it is to this particular weakness, and (3) the majority baselines might not be as indicative of "the difficulty of the task" (Bowman et al., 2015) as previously thought.

We are not the first to consider the inherent difficulty of NLI datasets. For example, MacCartney (2009) used a simple bag-of-words model to evaluate early iterations of Recognizing Textual Entailment (RTE) challenge sets.[1] Concerns have been raised previously about the hypotheses in the Stanford Natural Language Inference (SNLI) dataset specifically, such as by Rudinger et al. (2017) and in unpublished work.[2] Here, we survey of large number of existing NLI datasets under the lens of a hypothesis-only model.[3] Concurrently, Tsuchiya (2018) and Gururangan et al. (2018) similarly trained an NLI classifier with access limited to hypotheses and discovered similar results on three of the ten datasets that we study.

2 Motivation

Our approach is inspired by recent studies that show how biases in an NLU dataset allow models to perform well on the task without understanding the meaning of the text. In the Story Cloze task (Mostafazadeh et al., 2016, 2017), a model is presented with a short four-sentence narrative and asked to complete it by choosing one of two suggested concluding sentences. While the task is presented as a new common-sense reasoning framework, Schwartz et al. (2017b) achieved state-of-the-art performance by ignoring the narrative and training a linear classifier with features related to the writing style of the two potential endings, rather than their content. It has also been shown that features focusing on sentence length, sentiment, and negation are sufficient for achieving high accuracy on this dataset (Schwartz et al., 2017a; Cai et al., 2017; Bugert et al., 2017).

NLI is often viewed as an integral part of NLU. Condoravdi et al. (2003) argue that it is a necessary metric for evaluating an NLU system, since it forces a model to perform many distinct types of reasoning. Goldberg (2017) suggests that "solving [NLI] perfectly entails human level understanding of language", and Nangia et al. (2017) argue that "in order for a system to perform well at natural language inference, it needs to handle nearly the full complexity of natural language understanding." However, if biases in NLI datasets, especially those that do not reflect commonsense knowledge, allow models to achieve high levels of performance without needing to reason about hypotheses based on corresponding contexts, our current datasets may fall short of these goals.

3 Methodology

We modify Conneau et al. (2017)'s `InferSent` method to train a neural model to classify just the hypotheses. We choose `InferSent` because it performed competitively with the best-scoring systems on the Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015), while being representative of the types of neural architectures commonly used for NLI tasks. `InferSent` uses a BiLSTM encoder, and constructs a sentence representation by max-pooling over its hidden states. This sentence representation of a hypothesis is used as input to a MLP classifier to predict the NLI tag.

We preprocess each recast dataset using the NLTK tokenizer (Loper and Bird, 2002). Following Conneau et al. (2017), we map the resulting tokens to 300-dimensional GloVe vectors (Pennington et al., 2014) trained on 840 billion tokens from the Common Crawl, using the GloVe OOV vector for unknown words. We optimize via SGD, with an initial learning rate of 0.1, and decay rate of 0.99. We allow at most 20 epochs of training with optional early stopping according to the following policy: when the accuracy on the development set decreases, we divide the learning rate by 5 and stop training when learning rate is $< 10^{-5}$.

4 Datasets

We collect ten NLI datasets and categorize them into three distinct groups based on the methods by which they were constructed. Table 1 summarizes the different NLI datasets that our investigation considers.

[1] MacCartney (2009), Ch. 2.2: *"the RTE1 test suite is the hardest, while the RTE2 test suite is roughly 4% easier, and the RTE3 test suite is roughly 9% easier."*

[2] A course project constituting independent discovery of our observations on SNLI: https://leonidk.com/pdfs/cs224u.pdf

[3] Our code and data can be found at https://github.com/azpoliak/hypothesis-only-NLI.

Creation Protocol	Dataset	Size	Classes	Example Hypothesis
Recast	DPR	3.4K	2	*People raise dogs because dogs are afraid of thieves*
	SPR	150K	2	*The judge was aware of the dismissing*
	FN+	150K	2	*the irish are actually principling to come home*
Judged	ADD-1	5K	2	*A small child staring at a young horse and a pony*
	SCITAIL	25K	2	*Humans typically have 23 pairs of chromosomes*
	SICK	10K	3	*Pasta is being put into a dish by a woman*
	MPE	10K	3	*A man smoking a cigarette*
	JOCI	30K	3	*The flooring is a horizontal surface*
Elicited	SNLI	550K	3	*An animal is jumping to catch an object*
	MNLI	425K	3	*Kyoto has a kabuki troupe and so does Osaka*

Table 1: Basic statistics about the NLI datasets we consider. 'Size' refers to the total number of labeled premise-hypothesis pairs in each dataset (for datasets with $> 100K$ examples, numbers are rounded down to the nearest $25K$). The 'Creation Protocol' column indicates how the dataset was created. The 'Class' column reports the number of class labels/tags. The last column shows an example hypothesis from each dataset.

4.1 Human Elicited

In cases where humans were given a context and asked to generate a corresponding hypothesis and label, we consider these datasets to be **elicited**. Although we consider only two such datasets, they are the largest datasets included in our study and are currently popular amongst researchers. The elicited NLI datasets we look at are:

Stanford Natural Language Inference (SNLI) To create SNLI, Bowman et al. (2015) showed crowdsourced workers a premise sentence (sourced from Flickr image captions), and asked them to generate a corresponding hypothesis sentence for each of the three labels (ENTAILMENT, NEUTRAL, CONTRADICTION). SNLI is known to contain stereotypical biases based on gender, race, and ethnic stereotypes (Rudinger et al., 2017). Furthermore, Zhang et al. (2017) commented that this "elicitation protocols can lead to biased responses unlikely to contain a wide range of possible common-sense inferences."

Multi-NLI Multi-NLI is a recent expansion of SNLI aimed to add greater diversity to the existing dataset (Williams et al., 2017). Premises in Multi-NLI can originate from fictional stories, personal letters, telephone speech, and a 9/11 report.

4.2 Human Judged

Alternatively, if hypotheses and premises were automatically paired but *labeled* by a human, we consider the dataset to be **judged**. Our human-judged data sets are:

Sentences Involving Compositional Knowledge (SICK) To evaluate how well compositional distributional semantic models handle "challenging phenomena", Marelli et al. (2014) introduced SICK, which used rules to expand or normalize existing premises to create more difficult examples. Workers were asked to label the relatedness of these resulting pairs, and these labels were then converted into the same three-way label space as SNLI and Multi-NLI.

Add-one RTE This mixed-genre dataset tests whether NLI systems can understand adjective-noun compounds (Pavlick and Callison-Burch, 2016). Premise sentences were extracted from Annotated Gigaword (Napoles et al., 2012), image captions (Young et al., 2014), the Internet Argument Corpus (Walker et al., 2012), and fictional stories from the GutenTag dataset (Mac Kim and Cassidy, 2015). To create hypotheses, adjectives were removed or inserted before nouns in a premise, and crowd-sourced workers were asked to provide reliable labels (ENTAILED, NOT-ENTAILED).

SciTail Recently released, SciTail is an NLI dataset created from 4th grade science questions and multiple-choice answers (Khot et al., 2018). Hypotheses are assertions converted from question-answer pairs found in SciQ (Welbl et al., 2017). Hypotheses are automatically paired with premise sentences from domain specific texts (Clark et al., 2016), and labeled (ENTAILMENT, NEUTRAL) by crowd-sourced workers. Notably, the construction

method allows for the same sentence to appear as a hypothesis for more than one premise.

Multiple Premise Entailment (MPE) Unlike the other datasets we consider, the premises in MPE (Lai et al., 2017) are not single sentences, but four different captions that describe the same image in the FLICKR30K dataset (Plummer et al., 2015). Hypotheses were generated by simplifying either a fifth caption that describes the same image or a caption corresponding to a different image, and given the standard 3-way tags. Each hypothesis has at most a 50% overlap with the words in its corresponding premise. Since the hypotheses are still just one sentence, our hypothesis-only baseline can easily be applied to MPE.

Johns Hopkins Ordinal Common-Sense Inference (JOCI) JOCI labels context-hypothesis instances on an ordinal scale from *impossible* (1) to *very likely* (5) (Zhang et al., 2017). In JOCI, context (premise) sentences were taken from existing NLU datasets: SNLI, ROC Stories (Mostafazadeh et al., 2016), and COPA (Roemmele et al., 2011). Hypotheses were created automatically by systems trained to generate entailed facts from a premise.[4] Crowd-sourced workers labeled the likelihood of the hypothesis following from the premise on an *ordinal scale*. We convert these into a 3-way NLI tags where 1 maps to CONTRADICTION, 2-4 maps to NEUTRAL, and 5 maps to ENTAILMENT. Converting the annotations into a 3-way classification problem allows us to limit the range of the number of NLI label classes in our investigation.

4.3 Automatically Recast

If an NLI dataset was automatically generated from existing datasets for other NLP tasks, and sentence pairs were constructed and labeled with minimal human intervention, we refer to such a dataset as **recast**. We use the recast datasets from White et al. (2017):

Semantic Proto-Roles (SPR) Inspired by Dowty (1991)'s thematic role theory, Reisinger et al. (2015) introduced the Semantic Proto-Role (SPR) labeling task, which can be viewed as decomposing semantic roles into finer-grained properties, such as whether a predicate's argument was likely *aware* of the given predicated situation. 2-way labeled NLI sentence pairs were generated from SPR annotations by creating general templates.

Definite Pronoun Resolution (DPR) The DPR dataset targets an NLI model's ability to perform anaphora resolution (Rahman and Ng, 2012). In the original dataset, sentences contain two entities and one pronoun, and the task is to link the pronoun to its referent. In the recast version, the premises are the original sentences and the hypotheses are the same sentences with the pronoun replaced with its correct (ENTAILED) and incorrect (NOT-ENTAILED) referent. For example, *People raise dogs because they are obedient* and *People raise dogs because dogs are obedient* is such a context-hypothesis pair. We note that this mechanism would appear to maximally benefit a hypothesis-only approach, as the hypothesis semantically subsumes the context.

FrameNet Plus (FN+) Using paraphrases from PPDB (Ganitkevitch et al., 2013), Rastogi and Van Durme (2014) automatically replaced words with their paraphrases. Subsequently, Pavlick et al. (2015) asked crowd-source workers to judge how well a sentence with a paraphrase preserved the original sentence's meanings. In this NLI dataset that targets a model's ability to perform paraphrastic inference, premise sentences are the original sentences, the hypotheses are the edited versions, and the crowd-source judgments are converted to 2-way NLI-labels. For not-entailed examples, White et al. (2017) replaced a single token in a context sentence with a word that crowd-source workers labeled as not being a paraphrase of the token in the given context. In turn, we might suppose that positive entailments (1-b) are keeping in the spirit of NLI, but not-entailed examples might not because there are adequacy (1-c) and fluency (1-d) issues.[5]

(1) a. That is the way the system works
b. That is the way the framework works
c. That is the road the system works
d. That is the way the system creations

5 Results

Our goal is to determine whether a hypothesis-only model outperforms the majority baseline and investigate what may cause significant gains. In

[4] We only consider the hypotheses generated by either a seq2seq model or from external world knowledge.

[5] In these examples, (1-a) is the corresponding context.

Dataset	DEV Hyp-Only	DEV MAJ	DEV $\|\Delta\|$	DEV $\Delta\%$	TEST Hyp-Only	TEST MAJ	TEST $\|\Delta\|$	TEST $\Delta\%$	Baseline	SOTA
					Recast					
DPR	50.21	50.21	0.00	0.00	49.95	49.95	0.00	0.00	49.5	49.5
SPR	86.21	65.27	+20.94	+32.08	86.57	65.44	+21.13	+32.29	80.6	80.6
FN+	62.43	56.79	+5.64	+9.31	61.11	57.48	+3.63	+6.32	80.5	80.5
					Human Judged					
ADD-1	75.10	75.10	0.00	0.00	85.27	85.27	0.00	0.00	92.2	92.2
SciTail	66.56	50.38	+16.18	+32.12	66.56	60.04	+6.52	+10.86	70.6	77.3
SICK	56.76	56.76	0.00	0.00	56.87	56.87	0.00	0.00	56.87	84.6
MPE	40.20	40.20	0.00	0.00	42.40	42.40	0.00	0.00	41.7	56.3
JOCI	61.64	57.74	+3.90	+6.75	62.61	57.26	+5.35	+9.34	–	–
					Human Elicited					
SNLI	69.17	33.82	+35.35	+104.52	69.00	34.28	+34.72	+101.28	78.2	89.3
MNLI-1	55.52	35.45	+20.07	+56.61	–	35.6	– –		72.3	80.60
MNLI-2	55.18	35.22	+19.96	+56.67	–	36.5	–	–	72.1	83.21

Table 2: NLI accuracies on each dataset. Columns 'Hyp-Only' and 'MAJ' indicates the accuracy of the hypothesis-only model and the majority baseline. $|\Delta|$ and $\Delta\%$ indicate the absolute difference in percentage points and the percentage increase between the Hyp-Only and MAJ. Blue numbers indicate that the hypothesis-model outperforms MAJ. In the right-most section, 'Baseline' indicates the original baseline on the test when the dataset was released and 'SOTA' indicates current state-of-the-art results. MNLI-1 is the matched version and MNLI-2 is the mismatched for MNLI. The names of datasets are italicized if containing $\leq 10K$ labeled examples.

such cases a hypothesis-only model should be used as a stronger baseline instead of the majority class baseline. For all experiments except for JOCI, we use each NLI dataset's standard train, dev, and test splits.[6] Table 2 compares the hypothesis-only model's accuracy with the majority baseline on each dataset's dev and test set.[7]

Criticism of the Majority Baseline Across six of the ten datasets, our hypothesis-only model ***significantly outperforms*** the majority-baseline, even outperforming the best reported results on one dataset, recast SPR. This indicates that there exists a significant degree of exploitable signal that may help NLI models perform well on their corresponding test set without considering NLI contexts. From Table 2, it is unclear whether the construction method is responsible for these improvements. The largest relative gains are on human-elicited models where the hypothesis-only model more than doubles the majority baseline.

However, there are no obvious unifying trends across these datasets: Among the judged and recast datasets, where humans do not generate the NLI hypothesis, we observe lower performance margins between majority and hypothesis-only models compared to the elicited data sets. However, the baseline performances of these models are noticeably larger than on SNLI and Multi-NLI. 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. Therefore, we believe that models tested on SNLI and Multi-NLI should include a baseline version of the model that only accesses hypotheses.

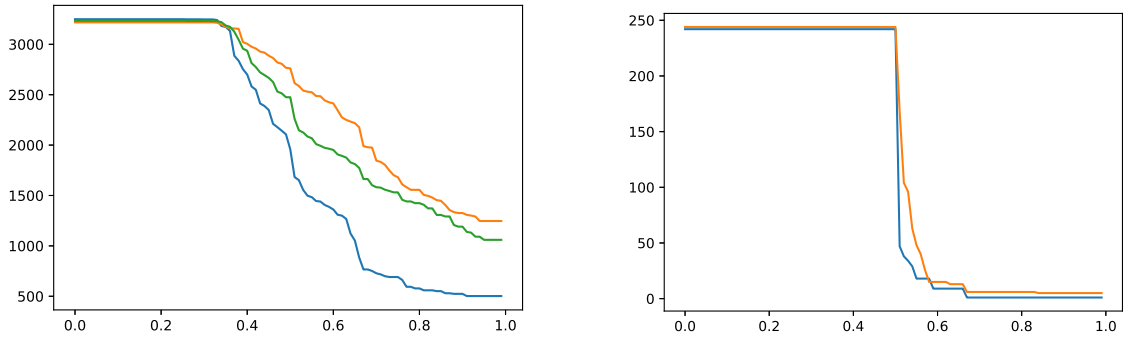
We do not observe general trends across the datasets based on their construction methodology. On three of the five human judged datasets, the hypothesis-only model defaults to labeling each instance with the majority class tag. We find the same behavior in one recast dataset (DPR). However, across both these categories we find smaller relative improvements than on SNLI and Multi-NLI. These results suggest the existence of exploitable signal in the datasets that is unrelated to NLI contexts. Our focus now shifts to identifying precisely what these signals might be and understanding why they may appear in NLI hypotheses.

6 Statistical Irregularities

We are interested in determining what characteristics in the datasets may be responsible for the hypothesis-only model often outperforming the majority baseline. Here, we investigate the importance of specific words, grammaticality, and lexical semantics.

[6] JOCI was not released with such splits so we randomly split the dataset into such a partition with 80:10:10 ratios.

[7] We only report results on the Multi-NLI development set since the test labels are only accessible on Kaggle.



(a) SNLI

(b) DPR

Figure 2: Plots showing the number of sentences per each label (Y-axis) that contain at least one word w such that $p(l|w) >= x$ for at least one label l. Colors indicate different labels. Intuitively, for a sliding definition of what value of $p(l|w)$ might constitute a "give-away" the Y-axis shows the proportion of sentences that can be trivially answered for each class.

6.1 Can Labels be Inferred from Single Words?

Since words in hypotheses have a distribution over the class of labels, we can determine the conditional probability of a label l given the word w by

$$p(l|w) = \frac{count(w,l)}{count(w)} \quad (1)$$

If $p(l|w)$ is highly skewed across labels, there exists the potential for a predictive bias. Consequently, such words may be "give-aways" that allow the hypothesis model to correctly predict an NLI label without considering the context.

If a single occurrence of a highly label-specific word would allow a sentence to be deterministically classified, how many sentences in a dataset are prone to being trivially labeled? The plots in Figure 2 answer this question for SNLI and DPR. The Y-value where $X = 1.0$ captures the number of such sentences. Other values of $X < 1.0$ can also have strong correlative effects, but a priori the relationship between the value of X and the coverage of trivially answerable instances in the data is unclear. We illustrate this relationship for varying values of $p(l|w)$. When $X = 0$, all words are considered highly-correlated with a specific class label, and thus the entire data set would be treated as trivially answerable.

In DPR, which has two class labels, because the uncertainty of a label is highest when $p(l|w) = 0.5$, the sharp drop as X deviates from this value indicates a weaker effect, where there are proportionally fewer sentences which contain highly label-specific words with respect to SNLI. As SNLI uses 3-way classification we see a gradual decline from 0.33.

6.2 What are "Give-away" Words?

Now that we analyzed the extent to which highly label-correlated words may exist across sentences in a given label, we would like to understand what these words are and why they exist.

Figure 3 reports some of the words with the highest $p(l|w)$ for SNLI, a human elicited dataset, and MPE, a human judged dataset, on which our hypothesis model performed identically to the majority baseline. Because many of the most discriminative words are low frequency, we report only words which occur at least five times. We rank the words according to their overall frequency, since this statistic is perhaps more indicative of a word w's effect on overall performance compared to $p(l|w)$ alone.

The score $p(l|w)$ of the words shown for SNLI deviate strongly, regardless of the label. In contrast, in MPE, scores are much closer to a uniform distribution of $p(l|w)$ across labels. Intuitively, the stronger the word's deviation, the stronger the potential for it to be a "give-away" word. A high word frequency indicates a greater potential of the word to affect the overall accuracy on NLI.

Qualitative Examples Turning our attention to the qualities of the words themselves, we can easily identify trends among the words used in contradictory hypotheses in SNLI. In our top-10 list, for example, three words refer to the act of sleeping. Upon inspecting corresponding context sentences, we find that many contexts, which are

SNLI

Word	Score	Freq
instrument	0.90	20
touching	0.83	12
least	0.90	10
Humans	0.88	8
transportation	0.86	7
speaking	0.86	7
screen	0.86	7
arts	0.86	7
activity	0.86	7
opposing	1.00	5

(a) entailment

Word	Score	Freq
tall	0.93	44
competition	0.88	24
because	0.83	23
birthday	0.85	20
mom	0.82	17
win	0.88	16
got	0.81	16
trip	0.93	15
tries	0.87	15
owner	0.87	15

(b) neutral

Word	Score	Freq
sleeping	0.88	108
driving	0.81	53
Nobody	1.00	52
alone	0.90	50
cat	0.84	49
asleep	0.91	43
no	0.84	31
empty	0.93	28
eats	0.83	24
sleeps	0.95	20

(c) contradiction

MPE

Word	Score	Freq
an	0.57	21
gathered	0.58	12
girl	0.50	12
trick	0.55	11
Dogs	0.55	11
watches	0.60	10
field	0.60	10
singing	0.50	10
outside	0.67	9
something	0.62	8

(d) entailment

Word	Score	Freq
smiling	0.56	16
An	0.60	10
for	0.56	9
front	0.75	8
camera	0.62	8
waiting	0.50	8
posing	0.50	8
Kids	0.57	7
smile	0.83	6
wall	0.50	6

(e) neutral

Word	Score	Freq
sitting	0.51	88
woman	0.55	80
men	0.56	34
Some	0.62	26
doing	0.59	22
Children	0.50	22
boy	0.67	21
having	0.65	20
sit	0.60	15
children	0.53	15

(f) contradiction

Figure 3: Lists of the most highly-correlated words in each dataset for given labels, thresholded to the top 10 and ranked according to frequency.

sourced from Flickr, naturally deal with activities. This leads us to believe that 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, i.e. describing the agent as sleeping. A second trend we notice is that universal negation constitutes four of the remaining seven terms in this list, and may also be used to similar effect.[8] The human-elicited protocol does not guide, nor incentivize crowd-source workers to come up with less obvious examples. If not properly controlled, elicited datasets may be prone to many label-specific terms. The existence of label-specific terms in human-elicited NLI datasets does not invalidate the datasets nor is surprising. Studies in eliciting norming data are prone to repeated responses across subjects (McRae et al., 2005) (see discussion in §2 of (Zhang et al., 2017)).

6.3 On the Role of Grammaticality

Like MPE, FN+ contains few high frequency words with high $p(l|w)$. However, unlike on MPE, our hypothesis-only model outperforms the majority-only baseline. If these gains do not arise from "give-away" words, then what is the statistical irregularity responsible for this discriminative power?

Upon further inspection, we notice an interesting imbalance in how our model performs for each of the two classes. The hypothesis-only model performs similarly to the majority baseline for entailed examples, while improving by over 34% those which are not entailed, as shown in Table 3.

As shown by White et al. (2017) and noticed

[8] These are "Nobody", "alone", "no", and "empty".

label	Hyp-Only	MAJ	Δ%
entailed	44.18	43.20	+2.27
not-entailed	76.31	56.79	+34.37

Table 3: Accuracies on FN+ for each class label.

by Poliak et al. (2018), FN+ contains more grammatical errors than the other recast datasets. We explore whether grammaticality could be the statistical irregularity exploited in this case. We manually sample a total of 200 FN+ sentences and categorize them based on their gold label and our model's prediction. Out of 50 sentences that the model correctly labeled as ENTAILED, 88% of them were grammatical. On the other-hand, 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. This suggests that a hypothesis-only model may be able to discover the correlation between grammaticality and NLI labels on this dataset.

6.4 Lexical Semantics

A survey of gains (Table 4) in the SPR dataset suggest a number of its property-driven hypotheses, such as *X was sentient in [the event]*, can be accurately guessed based on lexical semantics (background knowledge learned from training) of the argument. For example, the hypothesis-only baseline correctly predicts the truth of hypotheses in the dev set such as: *Experts were sentient ...* or *Mr. Falls was sentient ...*, and the falsity of *The campaign was sentient*, while failing on referring expressions like *Some* or *Each side*. A model exploiting regularities of the real world would seem to be a different category of dataset bias: while not strictly *wrong* from the perspective of NLU, one should be aware of what the hypothesis-only baseline is capable of, to recognize those cases where access to the context is required and therefore more interesting under NLI.

6.5 Open Questions

There may remain statistical irregularities, which we leave for future work to explore. For example, are there correlation between sentence length and label class in these data sets? Is there a particular construction method that minimizes the amount of "give-away" words present in the dataset? And lastly, our study is another in a line of research which looks for irregularities at the word level (MacCartney et al., 2008; MacCartney, 2009). Beyond bag-of-words, are there multi-word expressions or syntactic phenomena that might encode label biases?

Proto-Role	H-model	MAJ	Δ%
aware	88.70	59.94	+47.99
used in	77.30	52.72	+46.63
volitional	87.45	64.96	+34.62
physically existed	87.97	65.38	+34.56
caused	82.11	63.08	+30.18
sentient	94.35	76.26	+23.73
existed before	80.23	65.90	+21.75
changed	72.18	64.85	+11.29
chang. state	71.76	64.85	+10.65
existed after	79.29	72.91	+8.75
existed during	90.06	85.67	+5.13
location	93.83	91.21	+2.87
physical contact	89.33	86.92	+2.77
chang. possession	94.87	94.46	+0.44
moved	93.51	93.20	+0.34
stationary during	96.44	96.34	+0.11

Table 4: NLI accuracies on the SPR development data; each property appears in 956 hypotheses.

7 Related Work

Non-semantic information to help NLI In NLI datasets, non-semantic linguistic features have been used to improve NLI models. Vanderwende and Dolan (2006) and Blake (2007) demonstrate how sentence structure alone can provide a high signal for NLI. Instead of using external sources of knowledge, which was a common trend at the time, Blake (2007) improved results on RTE by combining syntactic features. More recently, Bar-Haim et al. (2015) introduce an inference formalism based on syntactic-parse trees.

World Knowledge and NLI As mentioned earlier, hypothesis-only models that perform without exploiting statistical irregularities may be performing NLI only in the sense that it is understanding language based on prior background knowledge. Here, we take the approach that *interesting* NLI should depend on both premise and hypotheses. Prior work in NLI reflect this approach. For example, Glickman and Dagan (2005) argue that "the notion of textual entailment is relevant only" for hypothesis that are not world facts, e.g. "Paris is the capital of France." Glickman et al. (2005a,b), introduce a probabilistic framework for NLI where the premise entails a hypothesis if, and only if, the probability of the hypothesis being true increases as a result of the premise.

NLI's resurgence Starting in the mid-2000's, multiple community-wide shared tasks focused on NLI, then commonly referred to as RTE, i.e, recognizing textual entailment. Starting with Dagan et al. (2006), there have been eight iterations of the PASCAL RTE challenge with the most recent being Dzikovska et al. (2013).[9] NLI datasets were relatively small, ranging from *thousands* to *tens of thousands* of labeled sentence pairs. In turn, NLI models often used alignment-based techniques (MacCartney et al., 2008) or manually engineered features (Androutsopoulos and Malakasiotis, 2010). Bowman et al. (2015) sparked a renewed interested in NLI, particularly among deep-learning researchers. By developing and releasing a large NLI dataset containing over $550K$ examples, Bowman et al. (2015) enabled the community to successfully apply deep learning models to the NLI problem.

8 Conclusion

We introduced a stronger baseline for ten NLI datasets. Our baseline reduces the task from labeling the relationship between two sentences to classifying a single hypothesis sentence. Our experiments demonstrated that in six of the ten datasets, always predicting the majority-class label is not a strong baseline, as it is significantly outperformed by the hypothesis-only model. Our analysis suggests that statistical irregularities, including word choice and grammaticality, may reduce the difficulty of the task on popular NLI datasets by not fully testing how well a model can determine whether the truth of a hypothesis follows from the truth of a corresponding premise.

We hope our findings will encourage the development of new NLI datasets which exhibit less exploitable irregularities, and that encourage the development of richer models of inference. As a baseline, new NLI models should be compared against a corresponding version that only accesses hypotheses. In future work, we plan to apply a similar hypothesis-only baseline to multi-modal tasks that attempt to challenge a system to understand and classify the relationship between two inputs, e.g. Visual QA (Antol et al., 2015).

[9]Technically Bentivogli et al. (2011) was the last challenge under PASCAL's aegis but Dzikovska et al. (2013) was branded as the 8th RTE challenge.

Acknowledgements

This work was supported by Johns Hopkins University, the Human Language Technology Center of Excellence (HLTCOE), DARPA LORELEI, and the NSF Graduate Research Fellowships Program (GRFP). We would also like to thank three anonymous reviewers for their feedback. 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

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Roy Bar-Haim, Ido Dagan, and Jonathan Berant. 2015. Knowledge-based textual inference via parse-tree transformations. *Journal of Artificial Intelligence Research*, 54:1–57.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2011. The seventh pascal recognizing textual entailment challenge.

Catherine Blake. 2007. The role of sentence structure in recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, pages 101–106, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Michael Bugert, Yevgeniy Puzikov, Andreas Rücklé, Judith Eckle-Kohler, Teresa Martin, Eugenio Martínez-Cámara, Daniil Sorokin, Maxime Peyrard, and Iryna Gurevych. 2017. Lsdsem 2017: Exploring data generation methods for the story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 56–61.

Zheng Cai, Lifu Tu, and Kevin Gimpel. 2017. Pay attention to the ending: Strong neural baselines for the

roc story cloze task. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 616–622.

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter D Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *AAAI*.

Cleo Condoravdi, Dick Crouch, Valeria De Paiva, Reinhard Stolle, and Daniel G Bobrow. 2003. Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 workshop on Text meaning-Volume 9*, pages 38–45. Association for Computational Linguistics.

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.

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.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, pages 547–619.

Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274, Atlanta, Georgia, USA. Association for Computational Linguistics.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764.

Oren Glickman and Ido Dagan. 2005. A probabilistic setting and lexical cooccurrence model for textual entailment. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 43–48. Association for Computational Linguistics.

Oren Glickman, Ido Dagan, and Moshe Koppel. 2005a. A probabilistic classification approach for lexical textual entailment. In *AAAI*.

Oren Glickman, Ido Dagan, and Moshe Koppel. 2005b. A probabilistic lexical approach to textual entailment. In *IJCAI*.

Yoav Goldberg. 2017. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proc. of NAACL*.

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

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.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pages 63–70. Association for Computational Linguistics.

Sunghwan Mac Kim and Steve Cassidy. 2015. Finding names in trove: Named entity recognition for australian historical newspapers. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 57–65.

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

Bill MacCartney, Michel Galley, and Christopher D Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of the conference on empirical methods in natural language processing*, pages 802–811. Association for Computational Linguistics.

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). ACL Anthology Identifier: L14-1314.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51, Valencia, Spain. Association for Computational Linguistics.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel Bowman. 2017. The repeval 2017 shared task: Multi-genre natural language inference with sentence representations. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 1–10.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, pages 95–100, Montréal, Canada. Association for Computational Linguistics.

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.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 2641–2649. IEEE.

Adam Poliak, Yonatan Belinkov, James Glass, and Benjamin Van Durme. 2018. On the evaluation of semantic phenomena in neural machine translation using natural language inference. In *Proceedings of the Annual Meeting of the North American Association of Computational Linguistics (NAACL)*.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea. Association for Computational Linguistics.

Pushpendre Rastogi and Benjamin Van Durme. 2014. Augmenting framenet via ppdb. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 1–5.

Drew Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. 2015. Semantic proto-roles. *Transactions of the Association for Computational Linguistics*, 3:475–488.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social Bias in Elicited Natural Language Inferences. In *The 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL): Workshop on Ethics in NLP*.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A Smith. 2017a. The effect of different writing tasks on linguistic style: A case study of the roc story cloze task. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 15–25.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A Smith. 2017b. Story cloze task: Uw nlp system. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 52–55.

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)*.

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.

Marilyn A Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 592–596. Association for Computational Linguistics.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106.

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.

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

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 for Computational Linguistics*, 5:379–395.