An Analysis of Natural Language Inference Benchmarks through the Lens of Negation

Md Mosharaf Hossain,[ϴ] **Venelin Kovatchev,**[ʒ] **Pranoy Dutta,**[ϴ] **Tiffany Kao,**[ϴ]
Elizabeth Wei,[ϴ] and **Eduardo Blanco**[ϴ]
[ϴ]University of North Texas [ʒ]University of Barcelona
mdmosharafhossain@my.unt.edu vkovatchev@ub.edu
{PranoyDutta,TiffanyKao,ElizabethWei}@my.unt.edu eduardo.blanco@unt.edu

Abstract

Negation is underrepresented in existing natural language inference benchmarks. Additionally, one can often ignore the few negations in existing benchmarks and still make the right inference judgments. In this paper, we present a new benchmark for natural language inference in which negation plays an important role. We also show that state-of-the-art transformers struggle making inference judgments with the new pairs.

1 Introduction

Natural language understanding remains an elusive goal except in limited scenarios. It is arguably the ultimate problem in natural language processing: to empower machines to understand language as generated by humans. The state of the art has seen tremendous progress in recent years, and has moved from symbolic representations (Bos et al., 2004; Artzi and Zettlemoyer, 2013) to distributional representations often learned from massive datasets (Devlin et al., 2019). Recognizing entailments (Dagan et al., 2006), identifying paraphrases (Das and Smith, 2009), determining semantic textual similarity (Agirre et al., 2012), and sentiment analysis (Pang and Lee, 2008) are but a few problems that require natural language understanding to a lesser or greater degree.

There are many benchmarks targeting the problems above, and they usually cast them as classification problems. A couple of popular evaluation platforms, GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019), aggregate benchmarks for some of the problems above and provide a single score for many tasks under the umbrella of natural language inference. State-of-the-art models are close to or even surpass human performance (Wang et al., 2019). This fact, however, is true only when evaluating models and humans with existing benchmarks. Indeed, researchers have pointed out weaknesses in benchmarks suggesting that we are evaluating models with examples that are much simpler than what humans are capable of (Section 3). Source text selection, annotation artifacts (Gururangan et al., 2018), and asking annotators—either experts or crowd workers—to write examples as opposed to retrieving real examples from previously generated language are a few of the culprits.

In this paper, we investigate the role of negation in a core natural language understanding task: natural language inference—in its most basic form, determining whether a *text* entails a *hypothesis*. Recognizing entailments has many applications including question answering (Trivedi et al., 2019), summarization (Pasunuru et al., 2017) and machine translation evaluation (Padó et al., 2009).

Negation relates an expression e to another expression with a meaning that is in some way opposed to the meaning of e (Horn and Wansing, 2017), thus it plays an important role in natural language understanding. Additionally, negation is ubiquitous in regular English texts: approximately 25% of English sentences contain negation depending on the domain and genre (Section 4). Despite these facts, negation is underrepresented and mostly irrelevant in existing benchmarks—one can literally disregard the negations and still make correct inference judgments in popular datasets. The work presented here addresses these shortcomings and makes the following contributions:[1]

1. We show that negation is underrepresented and often irrelevant in existing benchmarks.
2. We create new benchmarks for natural language inference in which negation plays a critical role to make inference judgments.
3. We demonstrate that state-of-the-art trans-

[1]New benchmarks and code available at https://github.com/mosharafhossain

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 9106–9118,
November 16–20, 2020.

formers trained with the original benchmarks are not robust when negation is present.

4. We provide empirical evidence that transformers may be unable to learn the intricacies of negation in the most challenging benchmark, which includes longer texts from many genres.

2 Background

The task of natural language inference or recognizing textual entailment consists in determining whether a *hypothesis* is true given a *text*. The original task considers two labels: entailment or no_entailment (Dagan et al., 2006), and a newer formulation considers three labels: entailment, contradiction or neutral (Giampiccolo et al., 2007). For example, the *text* "A person on a horse jumps over an airplane" entails *hypothesis* "A person is outdoors, on a horse," contradicts "A person is at a diner, ordering an omelette," and is neutral with respect to "A person is training his horse for a competition." We work with three existing benchmarks: a collection of RTE datasets (Dagan et al., 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009), SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018). The RTE datasets are smaller (5,767 text-hypothesis pairs) than SNLI and MNLI (569,033 and 431,997 pairs). MNLI is more challenging than RTE and SNLI: texts are longer and were selected from 10 genres including fiction and non-fiction as well as conversation transcripts. On the other hand, the texts in SNLI were selected from image captions. The hypotheses in SNLI and MNLI were crowdsourced, i.e., manually generated by non-experts.

Tables 2 and 4 show examples in the RTE, SNLI and MNLI benchmarks. We work with the formatted versions of these datasets in the GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) benchmarks for convenience.

3 Previous Work

Previous work has revealed weaknesses with the benchmarks we work with and that adversarial examples can break models for many natural language processing tasks. Adversarial examples consist of arguably trivial modifications to inputs that trick computational models. Some of them include misspellings (Pruthi et al., 2019), syntactically controlled paraphrases (Iyyer et al., 2018), lexical substitutions (Alzantot et al., 2018), and more elaborate substitutions (Ribeiro et al., 2018). More recently, Ribeiro et al. (2020) propose CHECKLIST, a task-agnostic strategy for testing NLP models. Their strategy can be used to identify which linguistic capabilities a model lacks. For example, they show that commercial systems for sentiment analysis are not robust when negation is present.

Regarding natural language inference, Poliak et al. (2018) show that models taking into account only hypotheses significantly outperform majority baselines, and Gururangan et al. (2018) discuss annotation artifacts, e.g., negation cues (not, never, etc.) are a strong indicator of contradictions. Glockner et al. (2018) show that models trained with SNLI fail to resolve new pairs that require simple lexical substitution, e.g., *holding a saxophone* contradicts *holding an electric guitar*. Naik et al. (2018) conclude that models are not robust to negation, but their only test is concatenating the tautology "and false is not true" to hypotheses. Wallace et al. (2019) introduce universal triggers and show that concatenating negation cues to SNLI hypotheses decreases accuracy to almost zero when the gold label is entailment or neutral.

The task of identifying paraphrases consists in determining whether two sentences have the same meaning, and can be casted—at least from a definitional perspective—as recognizing bidirectional entailments. Pruthi et al. (2019) show that computational models underperform in MRPC (Dolan et al., 2004) with adversarial misspellings, and Kovatchev et al. (2019) present a qualitative analysis of 11 state-of-the-art models (overall accuracies: 68–84%). When negation is present, however, accuracies drop to 33% (6 models) 67% (4 models) and 1% (1 model). Finally, Zhang et al. (2019) present a dataset for paraphrase identification including adversarial sentence pairs that are not paraphrases but have high word overlap. The new pairs helps training models robust to word scrambling.

The aforecited works do not investigate the role of negation in depth. Regarding paraphrase identification, previous work only has shown that models underperform with negation. Regarding natural language inference and negation, previous work considers negations only in the hypotheses—not the texts. Additionally, they only work with unrealistic negations that do not require models to do anything but ignore the negations. Indeed, they concatenate tokens including negations cues that are label-preserving and unrelated to the original texts and hypotheses. Unlike them, we (a) show

	#sents.	% w/ neg.
General English		
Online Reviews		
books	4,845,154	22.64
movies	616,287	28.97
Conversations		
oral	538,973	27.43
written	510,458	29.92
Wikipedia	2,735,930	8.69
Books	1,809,184	28.45
OntoNotes	63,918	17.14
NLI benchmarks		
RTE	16,389	7.16
SNLI	1,138,598	1.19
MNLI	883,436	22.63

Table 1: Percentage of sentences containing negation in general-purpose English corpora (reviews, conversations, Wikipedia, books and OntoNotes) and existing natural language inference benchmarks (also in English). Negation is underrepresented in RTE and SNLI.

that existing benchmarks do not properly account for negation in terms of frequency and difficulty, (b) create new benchmarks that require understanding negations, and (c) show that state-of-the-art models trained with existing corpora struggle with the new pairs including negation, and that the issue persists even if we fine-tune models with the new pairs in the most challenging benchmark, MNLI.

4 Negation in English and Natural Language Inference Benchmarks

Negation is pervasive in English (Morante and Sporleder, 2012), although there is limited empirical evidence from previous work (Councill et al., 2010; Elkin et al., 2005). In order to conduct a large-scale analysis and compare how often negation is present in English and existing natural language inference benchmarks, we employ a negation cue detector using a Bi-LSTM neural architecture with an additional CRF layer (Hossain et al., 2020). Trained and tested with CD-SCO, a corpus publicly available (Morante and Blanco, 2012), it obtains 0.92 F1. The supplemental materials provide more details regarding the architecture of the negation cue detector and the negation cues it detects.

Table 1 details the percentage of sentences with at least one negation in several large general-purpose English corpora. We work with online reviews (Wan et al., 2019; Maas et al., 2011), conversations (Chang et al., 2019), Wikipedia (50,000 pages with at least 20 views), 500 books from Project Gutenberg (Lahiri, 2014), and OntoNotes (Hovy et al., 2006) as released by Pradhan et al. (2011). The percentage of sentences containing negation is high: it ranges from 8.69% to 29.92% in all corpora, and is over 17% in all but Wikipedia. We note that negation is pervasive across domains and genres, including informal texts such as online reviews and both oral and written conversations (22.64–29.92%). Perhaps surprisingly, the percentage is very high in books (28.45%).

Table 1 also presents the percentage of sentences with negation in the three natural language inference benchmarks. Negation is clearly underrepresented in all of them except MNLI. These percentages do not invalidate the benchmarks. They show, however, that SNLI and RTE do not account for intricate linguistic phenomena such as negations. The reason for the low percentage in SNLI is that it uses texts from picture captions (Section 3), and captions describe pictures with affirmative statements (see examples in Tables 2 and 4).

The Role of Negation in Existing Natural Language Inference Benchmarks We conduct a manual qualitative analysis in order to (a) characterize the negations in RTE, SNLI and MNLI, and (b) assess how critical negation is to solve the few text-hypothesis pairs that include at least one negation in these benchmarks. We conduct the analysis with 100 text-hypothesis pairs containing negation from each benchmark (300 pairs total). From a linguistic perspective, most negations:

- are particles (no, not, n't, etc.) whose only function is to indicate negation (RTE: 62%, SNLI: 60%, MNLI: 84%),
- grammatically modify a verb (RTE: 62%, SNLI: 55%, MNLI: 81%), and
- scope over the main predicate (RTE: 52%, SNLI: 53%, MNLI: 62%).

These percentages are roughly uniformly distributed across labels.

In addition to looking at the negation cues in isolation, we also analyze the role of negation in making judgments. The first key distinction is whether dropping the negation changes the inference judgment (entailment or no_entailment in RTE; and entailment, neutral or contradiction in SNLI and MNLI). If it does not, we say the negation is ***unimportant*** (*important* otherwise). The

	Example
RTE	1) T: Mr Lopez Obrador, who lost July's presidential election by less than one percentage point, declared himself Mexico's "legitimate" president. H: Mr Lopez Obrador didn't loose the presidential election in July.
	2) T: If toxic waste containg cyanide is not disposed of properly, it may drain into ponds, streams, sewers, and reservoirs. H: Leaks into environment are caused by bad disposal of toxic waste containing cyanide.
	3) T: Toshiba has produced a fuel cell with no moving parts. H: Toshiba has no moving parts.
SNLI	4) T: A fighter jet plane is landed outside. H: The fighter jet plane is not moving.
	5) T: A man with no shirt on is performing with a baton. H: A man is doing things with a baton.
	6) T: A homeless man carries a sign that says "hungry". H: The man does not have a home.
MNLI	7) T: It was summertime the air conditioner was on the door was closed and i couldn't knock because i had to hold the jack with the other hand i finally with my elbow rang the doorbell and mother came to the door. H: The wintertime is when the air conditioning was on, I couldn't ring the doorbell because it was frozen.
	8) T: It runs advertisements for its supporters at the top of shows and strikes business deals with MCI, TCI, and Disney, but still insists it's not commercial. H: It runs ads for its supporters at shows and strikes business deals, but insists it is not commercial.

Table 2: Examples of the few text-hypothesis pairs that contain negation in the three natural language inference corpora we work with (RTE, SNLI and MNLI). Negation cues are underlined, and we have made minimal edits to some examples so that they fit within the width of the table.

second key distinction is whether the negation is ***aligned***, i.e., whether there is a semantic *alignment* between what is negated in the text (or hypothesis) and a chunk of the hypothesis (or text). We further identify ***negated alignments***, i.e., alignments in which the alignment is also *negated*.

Table 2 exemplifies this classification with the three benchmarks. Regarding SNLI, the negation in the hypothesis of Example (4) is important: *landed* entails *not moving*, at least according to the SNLI annotators, who were describing pictures thus (presumably) couldn't really tell if the plane was (a) completely stopped or taxiing after landing (and thus still moving). The negation in the text of Example (5), however, is unimportant: *A man with no shirt on is performing with a baton* entails *A man is doing things with a baton* regardless of whether the man has a shirt. Simply put, the negation plays no role in making the correct inference judgment. In Examples (4) and (6), the negations align but in Example (5), the negation does not align. Specifically, the alignments of the negations in the text and hypothesis of Example (6) are *negated*: *homeless* aligns with *does not have a home*, and both are negated. The alignment of the negation in the hypothesis of Pair (4), on the other hand, is not negated: *not moving* aligns with *landed*, and the latter is not negated. The categorization of the negations in text-hypothesis pairs from RTE and MNLI examples is as follows:

- RTE. The negation in the hypothesis of Example (1) is important, and it aligns but the alignment is not negated (*didn't loose* – lost). In Example (2), the negation in the text falls under the same categories: important and aligned, and the alignment is not negated (*not disposed of properly – bad disposal*). In Example (3), on the other hand, the negations are unimportant and aligned, in fact, there is an identical (and negated) alignment (*no moving parts* in both the text and hypothesis).
- MNLI. The negation in the text of Example (7), *I couldn't knock*, is unimportant and not aligned. Indeed, the first clause in both the text and hypothesis, which do not contain negation, are sufficient to solve the pair: *the air conditioning being on in wintertime* is not entailed by *the air conditioning being on in summertime*. The negation in the hypothesis of Example (7), however, is also unimportant but aligned (*I couldn't ring the doorbell – my elbow rang the doorbell*), although the alignment is not negated. This negation is unimportant for the same reason: one can make the correct inference judgment disregarding the negation altogether. The negations in Example (8) are similar to the ones in Example (3): unimportant and aligned, although this time the alignments are almost identical (*it's not commercial – it is not commercial*).

	RTE			SNLI				MNLI			
	E	¬E	All	E	C	N	All	E	C	N	All
% unimportant	77	75	76	38	24	93	48	78	24	83	52
% aligned	25	17	20	62	76	17	55	39	76	23	53
w/ negation	15	4	10	25	0	13	10	26	2	9	8

Table 3: Analysis of the few negations in the text-hypothesis pairs from the three natural language inference corpora we work with (RTE: 7.16% of pairs, SNLI: 1.19%, MNLI: 22.63%; Table 1). E stands for entailment, ¬E for no_entailment, C for contradiction and N for neutral. Many negations are unimportant, i.e., one can ignore them and still make the correct inference judgment.

Table 3 presents the analysis of the role of negation based on these categories. First, we note that one can often ignore negations without consequences: 76% of negations are unimportant in RTE, 48% in SNLI and 52% in MNLI. In RTE, negations are unimportant in text-hypothesis pairs regardless of the inference judgment (75-76%). In SNLI and MNLI, however, negations are almost always unimportant in neutral text-hypothesis pairs (93% in SNLI and 83% in MNLI), and they tend to be unimportant when the text entails the hypothesis (78% in MNLI and 38% in SNLI). Second, we note that few negations *align* in RTE (entailment: 25%, no_entailment: 17%), but about half of them align in SNLI and MNLI (55% and 53%). The percentage of aligned negations heavily depends on the inference judgment in SNLI and MNLI, and in RTE to a lesser degree (entailment is 50% more likely). More interestingly, whether the alignment is negated is a clear sign of the inference judgment. In RTE, the alignments are rarely negated in no_entailment pairs (4% overall, 23.5% of aligned pairs), but that is not the case with entailment pairs (15% overall, 60% of aligned pairs). In SNLI, the differences are larger: 40.3% of aligned pairs labeled *entailment* are negated. We observe a similar pattern in the negations from MNLI: alignments are rarely negated in contradictions (2.6% of aligned pairs), and most alignments are negated in entailment pairs (66.7% of aligned pairs).

5 A Benchmark for Natural Language Understanding with Negation

We create new benchmarks in which negation plays an important role for natural language inference. The starting points are the original benchmarks, more specifically, we selected at random 500 text-hypothesis pairs from RTE, SNLI and MNLI (1,500 text-hypothesis pairs total). We work with pairs from the training and development splits as GLUE and SuperGLUE do not include gold labels for some test splits. Then, we follow three steps for each of the selected original pairs. In the remaining of the paper, we use T and H to refer to texts and hypotheses in RTE, SNLI and MNLI.

1. Add negation manually to the main verb in T and H to obtain T_{neg} and H_{neg}.
2. Generate three new pairs automatically by combining the elements in the original pair (T and H) and the results of Step (1) (T_{neg} and H_{neg}). This results in the following pairs: T_{neg}-H, T-H_{neg} and T_{neg}-H_{neg}.
3. Manually annotate the pairs from Step (2) using the labels from the original benchmarks (RTE: entailment or no_entailment; SNLI and MNLI: entailment, contradiction or neutral).

These steps result in 4,500 new pairs and their judgments (3 per original pair, 1,500 from each RTE, SNLI and MNLI). Note that the negations are rather simple—adding *not* to the main verb, and adding auxiliaries and fixing verb tense if needed—but are realistic in the sense that the resulting texts and hypotheses follow proper English grammar. Additionally, the new pairs including negation are not more difficult than the original pairs except for the presence of negation. In particular, they do not require additional lexical inference and the overall topic described does not change.

Table 4 exemplifies the new pairs with negation. While negating the main verb (Step 1) is a relatively straightforward step, note that annotating the three new pairs including negation (Step 3) requires more attention from annotators. In other words, the inference judgment for the original T-H pair does not unequivocally indicate the inference judgment for the three new pairs that include negation. Indeed, the two examples generated from RTE in Table 4 show that when the original text entails the

	Original pair	New pair w/ negation
RTE	T: Tropical Storm Debby is blamed for several deaths across the Caribbean. H: A tropical storm has caused loss of life. *Judgments:* T-H: entailment, T_{neg}-H: no_entailment, T-H_{neg}: no_entailment, T_{neg}-H_{neg}: entailment	T_{neg}: Tropical Storm Debby is not blamed for several deaths across the Caribbean. H_{neg}: A tropical storm has not caused loss of life.
	T: Dr. Pridi was forced into exile, and Field Marshal Pibul again assumed power. H: Pibul was a field marshal. *Judgments:* T-H: entailment, T_{neg}-H: entailment, T-H_{neg}: no_entailment, T_{neg}-H_{neg}: no_entailment	T_{neg}: Dr. Pridi was not forced into exile, and Field Marshal Pibul again assumed power. H_{neg}: Pibul was not a field marshal.
SNLI	T: Two people are working on computers. H: Two people are near the computers. *Judgments:* T-H: entailment, T_{neg}-H: neutral, T-H_{neg}: contradiction, T_{neg}-H_{neg}: neutral	T_{neg}: Two people are not working on computers. H_{neg}: Two people are not near computers.
	T: Young man walking dog. H: The man is walking his cat. *Judgments:* T-H: contradiction, T_{neg}-H: neutral, T-H_{neg}: entailment, T_{neg}-H_{neg}: neutral	T_{neg}: Young man is not walking dog. H_{neg}: The man is not walking his cat.
MNLI	T: The lot upon which it is built had been vacant. H: The lot had been vacant. *Judgments:* T-H: entailment, T_{neg}-H: contradiction, T-H_{neg}: contradiction, T_{neg}-H_{neg}: entailment	T_{neg}: The lot upon which it is built had not been vacant. H_{neg}: The lot had not been vacant.
	T: Thursday's judge, the Honorable Charles Adams of the Coconino County Superior Court, agreed, but highly discouraged self-representation. H: Self-representation was encouraged by the Honorable Charles Adams. *Judgments:* T-H: contradiction, T_{neg}-H: contradiction, T-H_{neg}: entailment, T_{neg}-H_{neg}: entailment	T_{neg}: Thursday's judge, the Honorable Charles Adams of the Coconino County Superior Court, did not agree, but highly discouraged self-representation. H_{neg}: Self-representation was not encouraged by the Honorable Charles Adams.

Table 4: Examples of original pairs and new pairs generated after we manually introduce negation. Note that we (a) generate three new pairs after combining texts and hypotheses with and without negation (T-H is the original pair), and (b) manually annotate inference judgments for the three new pairs.

hypothesis, the three new text-hypothesis pairs may receive different inference judgments (in particular the judgments for T_{neg}-H and T_{neg}-H_{neg} are the opposite). The same is true across text-hypothesis pairs including negation and generated from different natural language inference benchmarks. For example, the text entails the hypothesis in the first examples shown from SNLI and MNLI, but the three new pairs including negation receive different judgments: neutral, contradiction and neutral; and contradiction, contradiction and entailment). The second examples created from SNLI and MNLI show the same phenomenon but with an original T-H pair labeled contradiction.

Annotation Process and Agreements. Three annotators and an additional adjudicator did the annotations described above in two phases.

In the first phase, the three annotators added negation to the main verbs of texts and hypotheses (Step 1). After a short training session, we decided to have only one annotator add negation in each original pair as the task is relatively straightforward. Any issues in this phase were detected during Phase 2. Text-hypothesis pairs with issues were discarded (only 5%) and additional pairs were collected to account for the discarded pairs (and still have 1,500 text-hypothesis pairs including negation and generated from each of the three benchmarks, 4,500 new text-hypothesis pairs in total).

In the second phase, the three annotators read the new pairs including negation (automatically generated in Step 2: T_{neg}-H, T-H_{neg} and T_{neg}-H_{neg}) and manually labeled them with inference judgments (Step 3). In this phase, each pair was annotated by two annotators independently, and the adjudicator resolved any disagreements. We calculated inter-annotator agreement prior to adjudication using Cohen's κ (Cohen, 1960). κ coefficients were 0.85 (RTE), 0.81 (SNLI) and 0.72 (MNLI). κ coefficients between 0.6 and 0.8 are considered substantial, and between 0.8 and 1.0 nearly perfect (Artstein and Poesio, 2008).

	RTE		SNLI			MNLI		
	%E	%¬E	%E	%C	%N	%E	%C	%N
T_{neg}-H	19.8	80.2	6.0	32.0	62.0	11.8	45.8	42.4
T-H_{neg}	9.0	91.0	21.4	41.0	37.6	24.0	47.6	28.4
T_{neg}-H_{neg}	34.4	65.6	22.2	8.0	69.8	38.6	14.4	47.0
All	21.1	78.9	16.5	27.0	56.5	24.8	35.9	39.3

Table 5: Label distribution in the new text-hypothesis pairs including negation depending on the source pairs they were generated from (RTE, SNLI or MNLI). Unlike the authors of the original benchmarks, we do not artificially force a uniform distribution. The pairs generated from MNLI, which are the longest and the only ones from many genres, are the most balanced (majority baseline accuracy: 39.3%).

	RTE			SNLI				MNLI			
	E	¬E	All	E	C	N	All	E	C	N	All
% unimportant	52	56	56	17	24	61	42	12	19	63	43
% aligned	76	60	62	78	76	42	59	84	84	42	61
w/ negation	52	10	17	28	0	2	5	48	3	7	14

Table 6: Analysis of the negations in the text-hypothesis pairs in the new benchmarks. E stands for entailment, ¬E for no_entailment, C for contradiction and N for neutral. Some negations are unimportant, but the percentage of important negations in the new text-hypothesis pairs is higher than those in the original corpora (Table 3).

Label Distributions. The original RTE, SNLI and MNLI benchmarks contain, by design, text-hypothesis pairs with roughly uniform judgment distributions. Thus, the majority baseline obtains roughly 50% accuracy in RTE (2 labels) and 33% in SNLI and MNLI (3 labels).

Our new benchmarks including negation do not have a uniform judgment distribution (Table 5), although the pairs generated from MNLI are close (entailment: 24.8%, contradiction: 35.9%, and neutral: 39.3%). We acknowledge that the label distribution in the new pairs generated from RTE (majority baseline: 78.9%) and, to a certain degree, SNLI (majority baseline: 56.5%) are not as challenging as the label distributions in the original pairs. As we shall see in Section 6, however, our experiments show that the ones from MNLI are a challenge for state-of-the-art transformers.

The Role of Negation. Table 6 presents the analysis of the role of negation in the new benchmarks using the categories presented in Section 4. We analyze 100 text-hypothesis pairs generated from each original benchmark (RTE, SNLI and MNLI). There are less unimportant negations in our new benchmarks than in the original corpora (Table 3). While many negations in the new pairs generated from RTE are unimportant (entailment: 52%, no_entailment: 56%), few negations in the pairs generated from SNLI and MNLI are unimportant, especially when the text entails or contradicts the hypothesis (SNLI: 17% and 24%, MNLI: 12% and 19%). Unsurprisingly, the percentage of aligned negations is higher in our corpus due to the steps we use to introduce negation, especially with in the new pairs generated from RTE (62% vs. 20%).

6 Experiments and Results

In order to assess whether state-of-the-art systems can solve the task of natural language inference when negation is present, we experiment with three state-of-the art transformers: BERT (Devlin et al., 2019), XLNet (Yang et al., 2019) and RoBERTa (Liu et al., 2019). We use the implementation and pretrained models by Wolf et al. (2019), and tune them to solve each benchmark. The supplemental materials provide details about (a) the hyperparameter settings we use to fine-tune these transformers, and (b) other implementation decisions.

We conduct two experiments. First, we assess whether these transformers tuned with the original train splits in RTE, SNLI and MNLI are capable of solving our new benchmarks including negation (Section 6.1). Second, we investigate if tuning with the new text-hypothesis pairs including negation improves the results (Section 6.2).

Test pairs	RTE				SNLI				MNLI			
	MB	[1]	[2]	[3]	MB	[1]	[2]	[3]	MB	[1]	[2]	[3]
Original												
dev	52.7	75.8	69.9	66.1	33.8	91.6	90.6	89.9	35.5	87.9	86.7	83.2
dev$_{neg}$	51.2	78.1	73.2	63.4	54.4	91.7	90.3	89.4	50.2	88.0	86.7	83.0
New w/ neg.												
T$_{neg}$-H	80.2	70.8	69.0	65.2	62.0	46.4	39.8	32.6	45.8	66.2	63.8	65.6
T-H$_{neg}$	91.0	51.4	44.2	39.2	41.0	63.6	67.4	58.8	47.6	70.4	69.8	62.4
T$_{neg}$-H$_{neg}$	65.6	65.4	69.6	68.4	69.8	45.8	47.2	41.8	47.0	63.6	65.4	63.6
All	78.9	62.5	60.9	57.6	56.5	51.9	51.5	44.4	39.3	66.7	66.3	63.9

Table 7: Results obtained with state-of-the-art models trained with the original training split for each benchmark and evaluated with (a) the original development split (dev), (b) pairs in the original development split containing negation (dev$_{neg}$), and (c) the new pairs containing negation. MB stands for the majority baseline, [1] for RoBERTa (Liu et al., 2019), [2] for XLNet (Yang et al., 2019) and [3] for BERT (Devlin et al., 2019).

Train pairs	RTE			SNLI			MNLI		
	[1]	[2]	[3]	[1]	[2]	[3]	[1]	[2]	[3]
Original	64.4	61.1	59.3	52.0	53.1	43.3	64.0	64.4	63.8
+ 70% new w/ neg.	88.2	87.3	83.8	75.3	74.2	69.1	67.3	70.4	66.4

Table 8: Results obtained testing with 30% of the new text-hypothesis pairs containing negation and training with either (a) the original train split from each benchmark or (b) the original train split from each benchmark and 70% of the new pairs containing negation. [1] stands for RoBERTa (Liu et al., 2019), [2] for XLNet (Yang et al., 2019) and [3] for BERT (Devlin et al., 2019). None of the transformers benefit from training with a portion of the pairs that include negation when tested with MNLI, which contains longer and more diverse text-hypothesis pairs.

6.1 Training with Existing Benchmarks

Can transformers solve the new text-hypothesis pairs including negation if trained with existing benchmarks? No, they cannot (Table 7). Indeed, the three transformers obtain worse results with the new pairs including negation, especially with SNLI (≈50% drop with the three transformers). These results might be unsurprising with SNLI and RTE since the original text-hypothesis pairs included few negations (1.19% and 7.16%, Table 1). The pattern is also true, however, with MNLI: we observe relative drops ranging from 23.0 to 24.2% despite 22.63% of text-hypothesis contain a negation in MNLI (Table 1). Comparing with the results obtained with the majority baseline, we observe that the transformers do not learn to solve pairs with negation unless they are tuned with pairs including negation (Section 6.2). Indeed, all of them obtain worse results than the majority baseline in RTE and SNLI, but not in MNLI.

We make a couple additional observations from the results in Table 7. First, the transformers solve the few text-hypothesis pairs including negation in the original benchmarks (dev$_{neg}$) as good (SNLI, MNLI) or better (RTE) than all pairs (dev). In other words, as our analysis of the role of negation in existing benchmarks points out (Section 4), negations do not bring additional complexity in these benchmarks. Second, RoBERTa and XLNet obtain roughly the same results with the new pairs including negation, but BERT falls slightly behind.

6.2 Fine-Tuning with New Pairs Containing Negation

Can transformers solve the new text-hypothesis pairs including negation if retrained with some of the new pairs including negation? Only to a certain degree: with SNLI, they benefit but underperform with respect to the original pairs; and with MNLI, they only benefit slightly.

In order to investigate whether the transformers can learn to make inference judgments when negation must be considered, we divide the new text-hypothesis pairs containing negation into training (70%) and test (30%) splits. Table 8 shows the results obtained with the new test split and the three

transformers trained with (a) the training split in the original benchmarks and (b) the training split in the original benchmarks combined with the training split with pairs containing negation. We observe that the transformers only learn to solve the new pairs including negation in the latter training scenario, but only partially. Indeed, we only observe a large improvement (59.3–64.4% vs. 83.8–88.2%) with the new pairs generated from RTE, which are also the only pairs that obtain higher accuracies than the original development split (83.8–88.2% vs. 66.1–75.8%). With the new pairs generated from SNLI, there is a substantial improvement after fine-tuning (43.3–53.1% vs. 69.1–75.3%) but the three transformers still obtain substantially worse results than with the original development split (69.1–75.3% vs. 89.9–91.6%). Finally, the transformers only benefit marginally from fine-tuning with the new pairs including negation and generated from MNLI (63.8–64.4% vs. 66.4–70.4%). Similar to the results obtained with pairs generated from SNLI, the transformers obtain substantially worse results than with the original development split in MNLI (66.4–70.4% vs. 83.2–87.9%). These results lead to the conclusion that natural language inference when negation is present remains an unsolved challenge.

7 Conclusions

Negation is ubiquitous in English and critical to understand language and make inferences, as it denies or inverts meaning. Despite these facts, negation is underrepresented in some natural language inference benchmarks (RTE and SNLI). Additionally, one can ignore negation and still make the correct inference judgment with many text-hypothesis pairs in existing natural language inference benchmarks (RTE, SNLI and MNLI).

In this paper, we have presented a new benchmark of text-hypothesis pairs containing negation (4,500 pairs). We generate and annotate these pairs after systematically adding negation to the main verb of the texts and hypotheses—either one or both—from RTE, SNLI and MNLI thus they are as difficult to solve as the original pairs except for the presence of negation. State-of-the art transformers trained with the original training splits from RTE, SNLI and MNLI obtain much worse results results with the new benchmark than with the original pairs—including the few original text-hypothesis pairs that do contain negation. In addition, our experimental results show that transformers struggle even after fine-tuning with new pairs containing negation.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1845757. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. Funding was also provided by the Spanish Ministry of Science, Innovation, and Universities Project PGC2018-096212-B-C33. The Titan Xp used for this research was donated by the NVIDIA Corporation. Computational resources were also provided by the UNT office of High-Performance Computing. We also thank the reviewers for insightful comments.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics*, 1:49–62.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the second PASCAL challenges workshop on recognising textual entailment*, volume 6, pages 6–4. Venice.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge.

Johan Bos, Stephen Clark, Mark Steedman, James R. Curran, and Julia Hockenmaier. 2004. Wide-coverage semantic representations from a CCG parser. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1240–1246, Geneva, Switzerland. COLING.

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*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2019. Convokit: The cornell conversational analysis toolkit.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Isaac Councill, Ryan McDonald, and Leonid Velikovich. 2010. What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 51–59, Uppsala, Sweden. University of Antwerp.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, pages 177–190, Berlin, Heidelberg. Springer-Verlag.

Dipanjan Das and Noah A. Smith. 2009. Paraphrase identification as probabilistic quasi-synchronous recognition. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 468–476, Suntec, Singapore. Association for Computational Linguistics.

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.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 350–356, Geneva, Switzerland. COLING.

Peter L Elkin, Steven H Brown, Brent A Bauer, Casey S Husser, William Carruth, Larry R Bergstrom, and Dietlind L Wahner-Roedler. 2005. A controlled trial of automated classification of negation from clinical notes. *BMC medical informatics and decision making*, 5(1):13.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics.

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

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7.

Laurence R. Horn and Heinrich Wansing. 2017. Negation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, spring 2017 edition. Metaphysics Research Lab, Stanford University.

Md Mosharaf Hossain, Kathleen Hamilton, Alexis Palmer, and Eduardo Blanco. 2020. Predicting the focus of negation: Model and error analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8389–8401, Online. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. 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 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Venelin Kovatchev, M. Antonia Marti, Maria Salamo, and Javier Beltran. 2019. A qualitative evaluation framework for paraphrase identification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 568–577, Varna, Bulgaria. INCOMA Ltd.

Shibamouli Lahiri. 2014. Complexity of Word Collocation Networks: A Preliminary Structural Analysis. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 96–105, Gothenburg, Sweden. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Roser Morante and Eduardo Blanco. 2012. *SEM 2012 shared task: Resolving the scope and focus of negation. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 265–274, Montréal, Canada. Association for Computational Linguistics.

Roser Morante and Walter Daelemans. 2012. ConanDoyle-neg: Annotation of negation cues and their scope in conan doyle stories. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1563–1568, Istanbul, Turkey. European Language Resources Association (ELRA).

Roser Morante and Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational Linguistics*, 38(2):223–260.

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

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Sebastian Padó, Michel Galley, Dan Jurafsky, and Christopher D. Manning. 2009. Robust machine translation evaluation with entailment features. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 297–305, Suntec, Singapore. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Ramakanth Pasunuru, Han Guo, and Mohit Bansal. 2017. Towards improving abstractive summarization via entailment generation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 27–32, Copenhagen, Denmark. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA. Association for Computational Linguistics.

Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5582–5591, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2017. Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. *arXiv preprint arXiv:1707.06799*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31.

Harsh Trivedi, Heeyoung Kwon, Tushar Khot, Ashish Sabharwal, and Niranjan Balasubramanian. 2019. Repurposing entailment for multi-hop question answering tasks. 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 2948–2958, Minneapolis, Minnesota. Association for Computational Linguistics.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian McAuley. 2019. Fine-grained spoiler detection from large-scale review corpora. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2605–2610, Florence, Italy. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32*, pages 3261–3275. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

A Identifying Negations

In order to identify negations in general English corpora as well as natural language inference corpora (RTE, SNLI, and MNLI, Section 4 in the paper), we develop a negation cue detector that consists of two-layer Bidirectional Long Short-Term Memory network with a Conditional Random Field layer (BiLSTM-CRF). This architecture (Figure 1) is similar to the one proposed by Reimers and Gurevych (2017). We train and evaluate the model with CD-SCO, a corpus of Conan Doyle stories with negation annotations (Morante and Daelemans, 2012; Morante and Blanco, 2012). CD-SCO includes common negation cues (e.g., never, no, n't), as well as prefixal (e.g., impossible, unbelievable) and suffixal negation (e.g., motionless).

We map each token in the input sentence to its 300-dimensional pre-trained GloVe embedding (Pennington et al., 2014). In addition, we extract token level universal POS tags using spaCy (Honnibal and Montani, 2017) and leverage another embedding (300-dimensional) to encode them. Embedding weights for universal POS are learned from scratch as part of the training of the network. We concatenate the word and POS embeddings, and feed them to the BILSTM-CRF architecture (size of cell state: 200 units). The learnt representations from the 2-layer BiLSTM are fed to a fully connected layer with ReLU activation function (Nair and Hinton, 2010). Finally, the CRF layer yields the final output.

We use the following labels to indicate whether a token is a negation cue: S_C (single-token negation cue, e.g., never, not), P_C (prefixal negation,

Hyperparameter	RTE			SNLI			MNLI		
	[1]	[2]	[3]	[1]	[2]	[3]	[1]	[2]	[3]
Batch size	16	8	8	32	32	32	32	32	32
Learning rate	2e-5	2e-5	2e-5	1e-5	1e-5	1e-5	2e-5	2e-5	2e-5
Epochs	10	50	50	3	3	3	3	3	3
Weight decay	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0

Table 9: Hyperparameters for fine-tuning the state-of-the-art systems on RTE, SNLI, and MNLI. [1] stands for RoBERTa (Liu et al., 2019), [2] for XLNet (Yang et al., 2019) and [3] for BERT (Devlin et al., 2019).

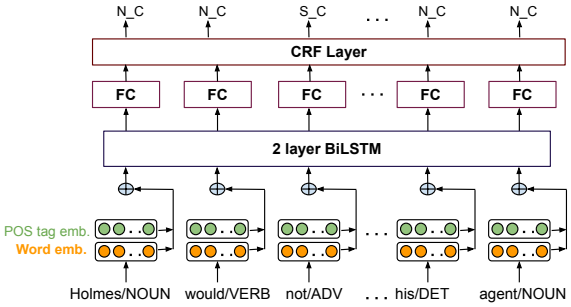

Figure 1: The BiLSTM-CRF architecture to identify negation cues. The input is a sentence. Each token is the concatenation of the word and its universal part-of-speech tag. The model outputs a sequence of labels indicating negation presence (S_C, P_C, SF_C or N_C). The example input sentence is "Holmes/NOUN would/VERB not/ADV listen/VERB to/ADP such/ADJ fancies/NOUN ,/PUNCT and/CCONJ I/PRON am/VERB his/DET agent/NOUN."

e.g., inconsistent), SF_C (suffixal negation, e.g., emotionless), and N_C (not a cue).

Training details. We merge the train and development instances from CD-SCO, and use 85% of the result as training and the remaining 15% as development. We evaluate our cue detector with the original test split from CD-SCO. We use the stochastic gradient descent algorithm with RMSProp optimizer (Tieleman and Hinton, 2012) for tuning weights. We set the batch size to 32, and the dropout and recurrent dropout are set to 30% for the LSTM layers. We stop the training process after the accuracy in the development split does not increase for 20 epochs, and the final model is the one which yields the highest accuracy in the development accuracy during the training process (not necessarily the model from the last epoch). Evaluating with the test set yields the following results: 92.75 Precision, 92.05 Recall, and 92.40 F1. While not perfect, the output of the cue detector is reliable, and an automatic detector is the only way to count negations in large corpora. The code is available at https://github.com/mosharafhossain/negation-cue.

The neural model has nearly 4.3 million parameters and takes 30 minutes on average to train on a CPU machine (Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz) with 64 GB of RAM.

B Fine-tuning Hyperparameters for State-of-the-Art Systems

For all the Transformer models, we set the maximum sequence length to 128. We use the Hugging Face implementation and pretrained models (Wolf et al., 2019). We work with the default settings for most of the hyperparameters except a few used to fine-tune to each benchmark. Table 9 shows the fine-tuned hyperparameters for the 3 transformers. Also, we use the base architectures for all the transformers (12-layer, 768-hidden, 12-heads).