An Analysis of Negation in Natural Language Understanding Corpora

Md Mosharaf Hossain,[ϴ] **Dhivya Chinnappa,**[ʊ] and **Eduardo Blanco**[ϒ]
[ϴ]Department of Computer Science and Engineering, University of North Texas
[ʊ]Thomson Reuters
[ϒ]School of Computing and Augmented Intelligence, Arizona State University
mdmosharafhossain@my.unt.edu dhivya.infant@gmail.com eduardo.blanco@asu.edu

Abstract

This paper analyzes negation in eight popular corpora spanning six natural language understanding tasks. We show that these corpora have few negations compared to general-purpose English, and that the few negations in them are often unimportant. Indeed, one can often ignore negations and still make the right predictions. Additionally, experimental results show that state-of-the-art transformers trained with these corpora obtain substantially worse results with instances that contain negation, especially if the negations are important. We conclude that new corpora accounting for negation are needed to solve natural language understanding tasks when negation is present.

1 Introduction

Natural language understanding (NLU) is an umbrella term used to refer to any task that requires text understanding. For example, question answering (Rajpurkar et al., 2016), information extraction (Stanovsky et al., 2018), coreference resolution (Wu et al., 2020), and machine reading (Yang et al., 2019), among many others, are tasks that fall under natural language understanding. The threshold for claiming that a system understands natural language is ever-moving. New corpora are often justified by pointing out that state-of-the-art models do not obtain good results. After years of steady improvements, more powerful models eventually obtain so-called human performance, and at that point new, more challenging corpora are created.

Many corpora for natural language understanding tasks contain language generated by annotators rather than retrieved from texts written independently of the corpus creation process. These corpora are certainly useful and have facilitated tremendous progress. Annotator-generated examples, however, carry the risk of evaluating systems with synthetic language that is not representative of language in the wild. For example, annotators are likely to use negation when asked to write a text that contradicts something despite contradictions in the wild need not have a negation (Gururangan et al., 2018). Recently, Kwiatkowski et al. (2019) present a large corpus for question answering that consists of natural questions (i.e., asked by somebody with a real information need) in order to encourage research in a more realistic scenario. This contrasts with previous corpora, where the questions were written by annotators after being told the answer (Rajpurkar et al., 2016).

In this paper, we explore the role of negation in eight corpora for six popular natural language understanding tasks. Our goal is to check whether negation plays the role it deserves in these tasks. To our surprise, we conclude that negation is virtually ignored by answering the following questions:[1]

1. Do NLU corpora contain as many negations as general-purpose texts? (they don't);
2. Do the (few) negations in NLU corpora play a role in solving the tasks? (they don't); and
3. Do state-of-the-art transformers trained with NLU corpora face challenges with instances that contain negation? (they do, especially if the negation is important).

2 Background and Related Work

We work with the eight corpora covering six tasks summarized below and exemplified in Table 2.

We select two corpora for question answering: CommonsenseQA (Talmor et al., 2019) and COPA (Roemmele et al., 2011). CommonsenseQA consists of multi-choice questions (5 candidate answers) that require some degree of commonsense. COPA presents a premise (e.g., *The man broke his toe*) and a question (e.g., *What was the cause of this?*) and the system must choose between two plausible alternatives (e.g. *He got a hole in his sock* or *He dropped a hammer on his foot*).

[1]Code and data available at https://github.com/mosharafhossain/negation-and-nlu.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 716 - 723

For textual similarity and paraphrasing, we select QQP[2] and STS-B (Cer et al., 2017). QQP consists of pairs of questions and the task is to determine whether they are paraphrases. STS-B consists of pairs of texts and the task is to determine how semantically similar they are with a score from 0 to 5.

We select one corpus for the remaining tasks. For inference, we work with QNLI (Rajpurkar et al., 2016), which consists in determining whether a text is a valid answer to a question. We use WiC (Pilehvar and Camacho-Collados, 2019) for word sense disambiguation. WiC consists in determining whether two instances of the same word (in two sentences; italicized in Table 2) are used with the same meaning. For coreference resolution, we choose WSC (Levesque et al., 2012), which consists in determining whether a pronoun and a noun phrase are co-referential (italicized in Table 2). Finally, we work with SST-2 (Socher et al., 2013) for sentiment analysis. The task consists in determining whether a sentence from a collection of movie reviews has positive or negative sentiment.

For convenience, we work with the formatted versions of these corpora in the GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) benchmarks. The only exception is CommonsenseQA, which is not part of these benchmarks.

Related Work Previous work has shown that SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) have annotation artifacts (e.g., negation is a strong indicator of *contradictions*) (Gururangan et al., 2018). The literature has also shown that simple adversarial attacks including negation cues are very effective (Naik et al., 2018; Wallace et al., 2019). Kovatchev et al. (2019) analyze 11 paraphrasing systems and show that they obtain substantially worse results when negation is present.

More recently, Ribeiro et al. (2020) show that negation is one of the linguistic phenomena commercial sentiment analysis struggle with. Several previous works have investigated the (lack of) ability of transformers to make inferences when negation is present. For example, Ettinger (2020) conclude that BERT is unable to complete sentences when negation is present. BERT also faces challenges solving the task of natural language inference (i.e., identifying entailments and contradictions) with monotonicity and negation (Geiger et al., 2020; Yanaka et al., 2019). Warstadt et al. (2019) show the limitations of BERT making acceptability judgments with sentences that contain negative polarity items. Most related to out work, Hossain et al. (2020) analyze the role of negation in three natural language inference corpora: RTE (Dagan et al., 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009), SNLI and MNLI. In this paper, we present a similar analysis, but we move beyond natural language inference and work with eight corpora spanning six natural language understanding tasks.

[2]https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs

	#sents.	% w/ neg.
Question Answering		
CommonsenseQA	12,102	14.5
COPA	1,000	0.8
Similarity and Paraphrasing		
QQP	1,590,482	8.1
STS-B	17,256	7.1
Inference		
QNLI	231,338	8.7
Word Sense Disambiguation		
WiC	14,932	8.2
Coreference Resolution		
WSC	804	26.2
Sentiment Analysis		
SST-2	70,042	16.0
General-purpose English		
all sentences	8,300,000	22.6–29.9
only questions	456,214	15.8–20.2

Table 1: Number of sentences and percentage of sentences containing negation in natural language understanding corpora. All but WSC contain substantially fewer negations than general-purpose English texts.

3 Research Questions and Analysis

Q1: Do natural language understanding corpora contain as many negations as general-purpose English texts? In order to automatically identify negation cues, we train a negation cue detector with the largest corpus available, ConanDoyle-neg (Morante and Daelemans, 2012). The cue detector is based on the RoBERTa pretrained language model (Liu et al., 2019); we provide details about the architecture and training process in Appendix A. Our cue detector obtains the best results to date: F1: 93.79 vs. 92.94 (Khandelwal and Sawant, 2020). ConanDoyle-neg (and thus our cue detector) identifies common negation cues such as *no*, *not*, *n't* and *never*, affixal negation cues such as *impossible* and *careless*, and lexical negations such as *deny* and *avoid*.

	Example		Important?
CmmsnsQA	[. . .] he (John) never saw the lady before. They were what? A) pay debts, B) slender, C) unacquainted, D) free flowing, E) sparse	C	✓
	When you travel you should what in case of unexpected costs? A) go somewhere, B) energy, C) spend frivilously, D) fly in airplane, E) have money	E	✗
QQP	What are some not-so-boring baby shower games ? What are some baby shower games that are actually fun?	yes	✓
	Who was philosophical guru of Shivaji Maharaj? What are the unknown facts of shivaji maharaj?	no	✗
STS-B	Colin Powell, the Secretary of State, said contacts with Iran would not stop. Secretary of State Colin Powell said yesterday that contacts with Iran would continue.	4.3	✓
	Well for one a being could have a non-physical existance and yet not even be in your mind. The difference is huge, as not all non-physical things exist in minds.	3.4	✗
QNLI	Who did BSkyB team up with as it was not part of consortium? While BSkyB had been excluded from being a part of the [. . .], BSkyB was able to join ITV Digital's free-to-air replacement, Freeview, in which it holds an equal stake [. . .]	yes	✓
	In what year did Lavoisier publish his work on combustion? In one experiment, Lavoisier observed that there was no overall increase in weight when tin and air were heated in a closed container.	no	✗
SST-2	It's not the ultimate depression-era gangster movie.	neg.	✓
	Whaley's determination to immerse you in sheer, unrelenting wretchedness is exhausting.	neg.	✗
WiC	The *intention* of this legislation is to boost the economy. Good *intentions* are not enough.	same	✗
WSC	*Sam and Amy* are passionately in love, but Amy's parents are unhappy about it, because *they* are only fifteen.	yes	✗

Table 2: Examples containing negation (underlined) from the validation datasets of the natural language understanding corpora we work with. The third column presents the expected answer for the example (a choice, judgment, or score depending on the task). The last column indicates whether the negation is important.

Table 1 presents the percentage of sentences that contain negation in (a) the eight corpora we work with and (b) general-purpose English. We take the latter percentage (all sentences) from Hossain et al. (2020), who run a negation cue detector in online reviews, conversations, and books. Additionally, we also present the percentages in questions. Negation is much less common in all natural language understanding corpora but WSC (0.8%–16%) than in general-purpose English (22.6%–29.9%). Note that negation is also underrepresented in corpora that primarily contain questions (general-purpose: 15.8%–20.2%; COPA: 0.8%, QQP: 8.1%).

Q2: Do the (few) negations in natural language understanding corpora play a role in solving the tasks? After showing that negation in underrepresented in natural language understanding corpora, we explore whether the few negations they contain are important. Given an instance from any of the corpora, we consider a negation *important* if removing it changes the ground truth. In other words, a negation is *unimportant* if one can ignore it and still solve the task at hand. Table 2 presents examples of important and unimportant negations.

We manually examine the negations in all instances containing negation from the validation split of each corpus except QQP, for which we examine 1,000 (out of 5,196). Note that COPA does not have any negations in the validation split, and many corpora have few instances containing negation (CommonsenseQA: 184, STS-B: 225, QNLI: 852, WiC: 99, WSC: 52, and SST-2: 263). We choose to work with the validation set because we want to compare results when negation is and is not important (Q3), and the ground truth for the test splits of some corpora are not publicly available.

We observe that (a) all negations in WiC and WSC are unimportant, and (b) the percentages of unimportant negations in CommonsenseQA, SST-2, QQP, STS-B, and QNLI are substantial: 45.1%, 63%, 97.4%, 95.6%, and 97.7%, respectively. These percentages indicate that one can safely ignore (almost) all negations and still solve the benchmarks. Despite the fact that negations are

		Example		Important?
CommonsenseQA	Syntactic	Where would a person live if they wanted no neighbors? A) housing estate, B) neighborhood, C) mars, D) woods, E) suburbs	D	✓
		The teacher doesn't tolerate noise during a test in their what? A) movie theatre, B) bowling alley, C) factory, D) store, E) classroom	E	✗
	Morpho.	What might result in an unsuccessful suicide attempt? A) die, B) interruption, C) bleed, D) hatred, E) dying	B	✓
		How are the conditions for someone who is living in a homeless shelter? A) sometimes bad, B) happy, C) respiration, D) growing older, E) death	A	✗
STS-2	Syntactic	Despite the evocative aesthetics evincing the hollow state of modern love life, the film never percolates beyond a monotonous whine.	neg.	✓
		Even if you don't think (kissinger's) any more guilty of criminal activity than most contemporary statesmen, he'd sure make a courtroom trial great fun to watch.	pos.	✗
	Morpho.	Makes for a pretty unpleasant viewing experience.	neg.	✓
		For anyone unfamiliar with pentacostal practices in general and theatrical phenomenon of hell houses in particular, it's an eye-opener .	pos.	✗

Table 3: Examples containing syntactic and morphological negation (underlined) from the validation datasets of CommonsenseQA and SST-2.

	CmmnsnsQA	COPA	QQP	STS-B	QNLI	WiC	WSC	SST-2
validation w/o neg	0.60	0.73	0.90	0.92 / 0.91	0.93	0.67	0.63	0.94
validation w/ neg	0.53	n/a	0.91	0.85 / 0.84	0.91	0.64	0.59	0.93
important (sample from Q2)	0.47	n/a	0.73	0.57 / 0.62	0.67	n/a	n/a	0.86
unimportant (sample from Q2)	0.62	n/a	0.92	0.85 / 0.84	0.92	0.64	0.59	0.95

Table 4: Results obtained with RoBERTa evaluating against (a) all instances with and without negation, and (b) the sample of instances with negation we analyze in detail (important and unimportant). Since the datasets are unbalanced, we report macro F1-score for all tasks except STS-B, for which we report Pearson and Spearman correlations. Results are slightly lower with negation, and substantially lower with *important* negations.

not important in WSC and WiC, they do affect the experimental results (details in Q3).

We also analyze the role of two major types of negation: syntactic (*not*, *no*, *never*, etc.) and morphological (i.e., affixes such as *un-*, *im-*, and *-less*). To this end, we work with CommonsenseQA and SST-2, which have lower percentages of unimportant negations (45.1% and 63%) than the other corpora we use (97.4%–100%). Table 3 provides examples of these two negation types. Perhaps unsurprisingly, syntactic negations are much more common than morphological negations (CommonsenseQA: 88.6% vs 11.4%, SST-2: 71.9% vs 28.1%). More importantly, syntactic negations are more often important in SST-2 (42.3% vs 23%), but both syntactic and morphological negation are roughly equaly important in CommonsenseQA (55.2% vs 52.4%).

Q3: Do state-of-the-art transformers trained with NLU corpora face challenges with instances that contain negation? We conduct experiments with RoBERTa (Liu et al., 2019). More specifically, we use the implementation by Phang et al. (2020) and train a model with the training split of each corpus. We refer the readers to the Appendix B for the details about these models and hyperparameters. We chose RoBERTa over other transformers because 4 out of the 10 best submissions to the SuperGLUE benchmark use it.[3]

Table 4 presents the results evaluating the models with the corresponding validation splits. RoBERTa obtains slightly worse results with the validation instances that have negation in all corpora; the only exception is QQP (F1: 0.90 vs. 0.91). These results lead to the conclusion that negation *may* only pose a small challenge to state-of-the-art transformers.

The results obtained evaluating with the important and unimportant negations from the samples analyzed in Question 2, however, provide a different picture. Indeed, we observe substantial drops in results in all tasks that have both kinds of negations. More specifically, we obtain 27% lower results

[3]https://super.gluebenchmark.com/leaderboard

with instances containing important negations in QNLI (F1: 0.92 vs. 0.67), 33%/26% lower in STS-B, 24% lower in CommonsenseQA, 21% lower in QQP, and 9% lower in SST. Further, even though all negations are unimportant in WiC and WSC, we observe a drop in performance for the instances with negation compared to the instances without negation (WiC: 0.64 vs 0.67 and WSC: 0.59 vs 0.63). We conclude that transformers trained with existing NLU corpora face challenges with instances that contain negation. These results raise two important questions for future research: Is negation an inherently challenging phenomenon for RoBERTa? How many instances with negation are required to solve a natural language understanding task?

4 Conclusions

We have analyzed the role of negation in eight natural language understanding corpora covering six tasks. Our analyses show that (a) all but WSC contain almost no negations or around 31%–54% of the negations found in general-purpose texts, (b) the few negations in these corpora are usually unimportant, and (c) RoBERTa obtains substantially worse results when negation is important.

Our analyses also provide some evidence that creating models to properly deal with negation may require both new corpora and more powerful models. The need for new corpora stems from the answers to Questions 1 and 2. The justification for powerful models is more subtle. We point out that the percentage of unimportant negations (Section 3) is only a weak indicator of the drop in results with important negations (Table 4). For example, we observe a 24% and 21% drop in results with important negations from CommonsenseQA and QQP despite 45% and 97% of negations are unimportant.

Negation reverses truth values thus solutions to any natural language understanding task should be robust when negation is present and important. To this end, our future work includes two lines of research. First, we plan to create benchmarks for the six tasks consisting of instances containing negation (50/50 split important/unimportant). Second, we plan to conduct probing experiments to investigate whether (and where) pretrained transformers capture the meaning of negation. Doing so may help us discover potential solutions to understand negation and make inferences.

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. 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. Further, we utilized computational resources from the Chameleon platform (Keahey et al., 2020). We also thank the reviewers for insightful comments.

References

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.

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.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

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.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models

partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173, Online. Association for Computational Linguistics.

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.

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.

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

Kate Keahey, Jason Anderson, Zhuo Zhen, Pierre Riteau, Paul Ruth, Dan Stanzione, Mert Cevik, Jacob Colleran, Haryadi S. Gunawi, Cody Hammock, Joe Mambretti, Alexander Barnes, François Halbach, Alex Rocha, and Joe Stubbs. 2020. Lessons learned from the chameleon testbed. In *Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC '20)*. USENIX Association.

Aditya Khandelwal and Suraj Sawant. 2020. NegBERT: A transfer learning approach for negation detection and scope resolution. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5739–5748, Marseille, France. European Language Resources Association.

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.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

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*.

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).

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.

Jason Phang, Phil Yeres, Jesse Swanson, Haokun Liu, Ian F. Tenney, Phu Mon Htut, Clara Vania, Alex Wang, and Samuel R. Bowman. 2020. `jiant` 2.0: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. 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.

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*, pages 90–95.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. 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 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, 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.

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.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. 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 2877–2887, Hong Kong, China. 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.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 250–255, Minneapolis, Minnesota. Association for Computational Linguistics.

An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. 2019. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2346–2357, Florence, Italy. Association for Computational Linguistics.

A Negation Cue Detection

We develop a negation cue detector (Section 3 in the paper) by utilizing the RoBERTa (base architecture; 12 layers) pre-trained model (Liu et al., 2019). We fine-tune the system on ConanDoyle-neg (Morante and Daelemans, 2012) corpus. While fine-training, the negation cues are marked with BIO (B: Beginning of cue, I: Inside of cue, O: Outside of cue) tagging scheme. The contextualized representations from the last layer of RoBERTa are passed to a fully connected (FC) layer. Finally, a conditional random field (CRF) layer produces the output sequence for the labels.

Our model yields the following results on the test set: 93.26 Precision, 94.32 Recall, and 93.79 F1. The neural model takes about two hours on average to train on a single GPU of NVIDIA Tesla K80. A list of the tuned hyperparameters that the model requires to achieve the above results is provided in Table 5. The code is available at `https://github.com/mosharafhossain/negation-and-nlu`.

Hyperparameter	
Max Epochs	50
Batch Size	10
Learning Rate (RoBERTa)	1e-5
Learning Rate (FC, CRF)	1e-3
Weight Decay (RoBERTa)	0.00001
Weight Decay (FC)	0.001
Grad Clipping	5.0
Warmup Epochs	5
Patience	15
Dropout	0.5

Table 5: Hyperparameters used to fine-tune the cue detector with ConanDoyle-neg (Morante and Daelemans, 2012) corpus. FC and CRF refers to fully connected and conditional random field layers, respectively.

	Hp-1	Hp-2	Hp-3
CmmnsnsQA	10	16	1e-5
COPA	50	16	1e-5
QQP	3	16	1e-5
STS-B	10	16	1e-5
QNLI	3	8	1e-5
WiC	10	16	1e-5
WSC	200	16	1e-6
SST-2	3	16	1e-5

Table 6: Hyperparameters used to fine-tune RoBERTa individually for each corpus. Hp-1, Hp-2, and Hp-3 refer to the number of epochs, batch size, and learning rate used in the training procedure. We use default settings for the other hyperparameters when we use the implementation by Phang et al. (2020).

B Hyperparameters to Fine-tune the System for Each of the NLU Tasks

We use an implementation by Phang et al. (2020) and fine-tune RoBERTa (base architecture; 12 layers) (Liu et al., 2019) model separately for each of the eight corpora. We use the default settings of the hyperparameters, except for a few, when fine-tuning the model on each benchmark. Table 6 shows tuned hyperparameters for each benchmark.