Leveraging Affirmative Interpretations from Negation Improves Natural Language Understanding

Md Mosharaf Hossain[ɵ,ʊ*] and **Eduardo Blanco**[ɤ]
[ɵ]Department of Computer Science and Engineering, University of North Texas
[ʊ]Amazon
[ɤ]Department of Computer Science, University of Arizona
mdmosharafhossain@my.unt.edu hosmdmos@amazon.com eduardoblanco@arizona.edu

Abstract

Negation poses a challenge in many natural language understanding tasks. Inspired by the fact that understanding a negated statement often requires humans to infer affirmative interpretations, in this paper we show that doing so benefits models for three natural language understanding tasks. We present an automated procedure to collect pairs of sentences with negation and their affirmative interpretations, resulting in over 150,000 pairs. Experimental results show that leveraging these pairs helps (a) T5 generate affirmative interpretations from negations in a previous benchmark, and (b) a RoBERTa-based classifier solve the task of natural language inference. We also leverage our pairs to build a plug-and-play neural generator that given a negated statement generates an affirmative interpretation. Then, we incorporate the pretrained generator into a RoBERTa-based classifier for sentiment analysis and show that doing so improves the results. Crucially, our proposal does not require any manual effort.

1 Introduction

Natural Language Understanding is a crucial component to build intelligent systems that interact with humans seamlessly. While recent papers sometimes report so-called superhuman performance, simple adversarial attacks including adding negation and other input modifications remain a challenge despite they are obvious to humans (Naik et al., 2018; Wallace et al., 2019). Further, many researchers have found that state-of-the-art systems struggle with texts containing negation. For example, Kassner and Schütze (2020) show that pretrained language models such as BERT (Devlin et al., 2019) do not differentiate between negated and non-negated cloze questions (e.g., *Birds cannot [MASK]* vs. *Birds can [MASK]*). Other studies show that transformers perform much worse in many other natural language understanding tasks when there is a negation in the input sentence (Ribeiro et al., 2020; Ettinger, 2020; Hossain et al., 2020b; Hosseini et al., 2021; Hossain et al., 2022a; Truong et al., 2022).

English-Norwegian (en-no) parallel sentences:
(en) *There is no more than one Truth.*
(no) *Og det finnes kun en Sannhet.*
Backtranslation: *And there is only one truth.*

English-Spanish (en-es) parallel sentences:
(en) *The term gained traction only after 1999.*
(es) *El término no se popularizó hasta después del 1999.*
Backtranslation: *The term was not popular until 1999.*

Figure 1: Parallel sentences from bitext corpora (English-Norwegian and English-Spanish) and backtranslations into English. Either the original English sentence or the backtranslation contains a negation, and the other one is an affirmative interpretation. In this paper, we show that leveraging sentences with negation and their affirmative interpretations is beneficial for several natural language understanding tasks including natural language inference and sentiment analysis.

In this paper, we address this challenge building upon the following observation: negation often carries affirmative meanings (Horn, 1989; Hasson and Glucksberg, 2006). For example, people intuitively understand that *John read part of the book* from *John didn't read the whole book*. Our fundamental idea is to leverage a large collection of sentences containing negation and their affirmative interpretations. We define an affirmative interpretation as a semantically equivalent sentence that does not contain negation. We explore this idea by automatically collecting pairs of sentences with negation and their affirmative interpretations from parallel corpora and backtranslating. Figure 1 exemplifies the idea with English-Norwegian and English-Spanish parallel sentences. Note that (a) either the original English sentence or the backtranslation have a negation (the one that does not is the affirmative interpretation) and (b) the meaning of both

*Work was done prior to joining Amazon.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5833 - 5847
December 7-11, 2022

is equivalent.

Armed with the large collection of sentences containing negation and their affirmative interpretations, we show that leveraging them yields improvements in three natural language understanding tasks. First, we address the problem of generating affirmatively interpretations in the AFIN benchmark (Hossain et al., 2022b), a collection of sentences with negation and their manually curated affirmative interpretations. Second, we address natural language inference using three common benchmarks: RTE (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). Third, we address sentiment analysis using SST-2 (Socher et al., 2013). The main contributions of this paper are:[1]

1. A large collection (153,273) of pairs of sentences containing negation and their affirmative interpretations. We present an automated procedure to get these pairs and an analysis of the negation types (single tokens, morphological, lexicalized, etc.).
2. Experimental results with the T5 transformer (Raffel et al., 2020) showing that blending our pairs during the fine-tuning process is beneficial to generate affirmative interpretations from the negations in AFIN.
3. Experimental results showing that a RoBERTa-based classifier (Liu et al., 2019) to solve the task of natural language inference benefits from training with new premise-hypothesis derived from our pairs (two entailments per pair).
4. Experimental results showing that a RoBERTa-based classifier for sentiment analysis benefits from a novel component that automatically generates affirmative interpretations from the input sentence.

The key resource enabling the experimental results is our large collection of pairs of sentences containing negation and their affirmative interpretations. As we shall see, the experiments under (2) and (3) are a somewhat straightforward applications of these pairs. The affirmative interpretation generator we use to improve sentiment analysis, however, has the potential to improve many natural language understanding tasks.

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

2 Related Work

Solving natural language understanding tasks when the input text contains negation is challenging. Researchers have approached negation processing mainly by identifying the scope (Vincze et al., 2008; Morante and Daelemans, 2012a) and focus (Blanco and Moldovan, 2011). Scope refers to the part of the meaning that is negated and focus refers to the part of the scope that is most prominently negated (Huddleston and Pullum, 2002). There are many works targeting scope detection (Fancellu et al., 2016, 2017; Li and Lu, 2018; Jumelet and Hupkes, 2018; Chen, 2019; Zhao and Bethard, 2020) and focus detection (Zou et al., 2014, 2015; Shen et al., 2019; Hossain et al., 2020a). While scope and focus pinpoint what is and what is not negated, they do not reveal affirmative interpretations as defined in this paper. Additionally, there is limited empirical evidence showing that scope or focus is beneficial to solve a natural language understanding task. Jiménez-Zafra et al. (2021) show that scope improves sentiment analysis, but they do not experiment with modern networks that may not benefit from explicit scope information.

Outside of scope and focus, Jiang et al. (2021) work with commonsense implications involving negations (e.g., "If X doesn't wear a mask" then "X is seen as carefree"). Closer to our work, Hosseini et al. (2021) pretrain BERT with an unlikelihood loss calculated with automatically obtained negated statements. Their negated statements do not preserve meaning. The authors show that their method, BERTNOT, outperforms BERT with LAMA (Kassner and Schütze, 2020) and the same natural language inference corpora we work with. The work proposed here outperforms theirs (Section 4.2) and does not require any manual effort.

We are not the first to work with affirmative interpretations from negated statements. For example, Sarabi et al. (2019) create a small corpus of verbal negations retrieved from Simple Wikipedia and their affirmative interpretations (total: 5,900). Simple Wikipedia is a version of Wikipedia that uses shorter sentences and simpler language. Hossain et al. (2022b) propose a question-answer driven approach to create AFIN, a collection of 3,001 sentences with negation and their affirmative interpretations. Both of these previous efforts employ humans to collect affirmative interpretations and neither one conducts extrinsic evaluations. Unlike

	Source	#parl. sents.	#pairs	%pairs
en-no	WikiMatrix	530,000	10,274	1.94
	CCMatrix	8,000,000	73,394	0.92
en-es	UNPC	2,800,000	28,028	1.00
	WikiMatrix	3,290,000	41,577	1.26
	All	14,620,000	153,273	1.05

Table 1: Number of parallel sentences in the English-Norwegian and English-Spanish parallel corpora we work with, and pairs of sentences with negation and affirmative interpretations we automatically generate via backtranslation. The yield (%pairs) is low, but as we shall see these pairs are useful to solve natural language understanding tasks when negation is present without hurting results when negation is not present.

them, we automatically collect pairs of sentences with negation and their affirmative interpretations. Additionally, extrinsic evaluations show that despite our collection procedure is noisy, leveraging our pairs is beneficial to solve three natural language understanding tasks.

3 Collecting Sentences with Negation and Their Affirmative Interpretations

This section outlines our approach to create a large collection of sentences containing negation and their affirmative interpretations. First, we present the sources of parallel corpora we work with. Second, we describe our multilingual negation cue detector to identify negation cues in the parallel sentences. Third, we describe the backtranslation step and a few checks to improve quality. Lastly, we present an analysis of the resulting sentences with negation and their affirmative interpretations.

3.1 Selecting Parallel Corpora

We select parallel sentences in English and either Norwegian or Spanish for two reasons: (a) large parallel corpora are available in these language pairs and (b) negation cue annotations are available in monolingual corpora for the three languages. The latter is a requirement to build a multilingual cue detector (Section 3.2). We extract the parallel sentences from three parallel corpora available in the OPUS portal (Tiedemann, 2012)): WikiMatrix (Schwenk et al., 2021a), CCMatrix (Schwenk et al., 2021b; Fan et al., 2021), and UNPC (Ziemski et al., 2016). Table 1 (Column 3) shows the number of parallel sentences we collect from each of the corpora and language pair (total: 14.6 million).

3.2 Identifying Negation Cues in Multiple Languages

In order to detect negation in the parallel sentences, we develop a multilingual negation cue detector that works with English, Norwegian, and Spanish texts. To this end, we fine-tune a multilingual BERT (mBERT)[2] (Devlin et al., 2019) with negation cue annotations in the three languages we work with: English (Morante and Daelemans, 2012b), Norwegian (Mæhlum et al., 2021), and Spanish (Jiménez-Zafra et al., 2018). We fine-tune jointly for all three languages by combining the original training splits into a multilingual training split. We terminate the training process after the F1 score in the (combined) development split does not increase for 5 epochs; the final model is the one which yields the highest F1 score during the training process. Additional details regarding training procedure and hyperparameters are provided in Appendix A. Our multilingual detector is not perfect but obtains competitive results (F1 scores): English: 91.96 (test split), Norwegian: 93.40 (test split), and Spanish: 84.41 (dev split, as gold annotations for the test split are not publicly available). The system detects various negation cue types including single tokens (no, never, etc.), affixal, and lexicalized negations (Section 3.4).

We use our multilingual cue detector to detect negation in the 14.6 million of parallel sentences. In the English-Norwegian parallel sentences (8.5M), negation is present in both sentences (WikiMatrix: 7.3%, CCMatrix: 14.2%), either sentence (WikiMatrix: 5.2%, CCMatrix: 5.2%), or neither sentence (WikiMatrix: 87.5%, CCMatrix: 80.6%). Similarly, in English-Spanish parallel sentences, negation is present in both sentences (UNPC: 10.7%, WikiMatrix: 5.7%), either sentence (UNPC: 4.6%, WikiMatrix: 4.4%), or neither sentence (UNPC: 84.7%, WikiMatrix: 89.9%). Since we are interested in sentences containing negation and their affirmative interpretations, we only keep the sentences in which either the source or target sentence contains negation.

3.3 Generating Affirmative Interpretations

After identifying negation cues in the parallel sentences, we backtranslate into English the sentence in the target language (either Norwegian or Spanish; they may or may not contain a negation). In

[2] https://github.com/google-research/bert/blob/master/multilingual.md

Negation Type	Examples
Single tokens (49.6%) *Cues: not, n't, no, never, without, nothing, nowhere, nobody, none, etc.*	They are still not integrated into the German community. They have yet to integrate into German society.
	I have no doubt that we will reach our goal. We shall surely get there!
	This process allows for higher precision that could never be achieved by hand. This process allows more precision than anyone performed manually.
Affixal (30.15%) *Cues: un-, in-, -less, etc.*	The north wing was left largely untouched and forms the present house. Only the North wing remained quite intact, and constitutes the current house.
	We fall in love, and any attempt at logic is useless. We fall in love and any attempt at logic is futile.
Lexicalized (8.76%) *Cues: prevent, lack, etc.*	A further problem was the lack of skilled labour. Another problem was the issue of obtaining sufficiently qualified personnel.
Multitoken (2.58%) *Cues: no longer, not at all, etc.*	After some time, the drainage of water no longer occurs. After a certain time, the drainage of water ends.
Multiple negations (8.95%)	The declaration before the courts is not valid if the child is not 14 years old. Any statement in a court is invalid if the child is below 14 years of age.

Table 2: Examples of sentences with negation and their affirmative interpretations automatically obtained from bitext corpora via backtranslation. We present examples for several negation types; common *single-tokens* that are not lexicalized negations are the most frequent. These sentences with negations and their affirmative interpretations come from our collection (Section 3) and include errors. For example, the affirmative interpretation in the second to last example includes *ends*, a lexicalized negation, because our cue detector did not identify it.

particular, we utilize Google Cloud Translation API.[3] Before backtranslating, we exclude sentences in the target language if they are longer than 40 tokens, as longer sentences tend to result in lower translation quality (Fonteyne et al., 2020).

Backtranslating into English from either Norwegian or Spanish may introduce or remove a negation cue. We discard such backtranslations since our goal is to obtain pairs of sentences containing negation and its affirmative interpretation (i.e., a semantically equivalent sentence that does not contain negation). The last two columns in Table 1 present how many pairs we obtain (total: 153,273). While the yield is small (1.05%), we note that the process is automated and could be expanded to use additional parallel corpora.

3.4 Quality and Analysis

The process to collect pairs of sentences with negation and their affirmative interpretations is noisy. First, the negation cue detector is not perfect thus there are pairs in which the affirmative interpretation contains a negation. Additionally, backtranslating introduces errors thus the affirmative interpretations are not always semantically equivalent to the sentences containing negation. Our goal is not to create a gold standard but to collect a large collection that we can leverage to improve models for natural language understanding tasks (Section 4).

Despite 100% correctness is not the goal, we conducted a manual validation with a random sample of 100 pairs. We discovered that 78% are correct, where correct means that the affirmative interpretation satisfies the definition (i.e., no negation and semantically equivalent to the sentence with negation). We found two main reasons for incorrect pairs. First, the negation cue detector sometimes fails to detect cues, resulting in affirmative interpretations that contain negation (e.g., *The execution had been unlawful*: *This act would have been illegal*; the prefix *il-* is not identified as a negation cue). Second, the backtranslation sometimes results in the original sentence without the negation cue and thus opposite meanings. (e.g., English-Norwegian parallel sentences: *How can you not enjoy this trip!*, *Hvordan kan du nyte denne turen!*; backtranslation: *How can you enjoy this trip!*).

Analyzing the Negations and Affirmative Interpretations The negation cue detector identifies

[3]Google Translate API - https://cloud.google.com/translate

several types of negation cues. As a result, our collection of pairs of sentences with negation and their affirmative interpretations includes several negation types (Table 2). Note that this table presents real examples from our collection including erroneous ones (e.g., some affirmative interpretations contain negation). The most frequent negation type (49.6%) are common single-token negation cues such as *not*, *n't* and *never*. Affixal negations are surprisingly common (30.15%) and include both prefixes (e.g., *untouched*) and suffixes (*useless*). Lexicalized negations usually take the form of a noun (e.g., lack, dismissal) or verb (e.g., prevent, avoid) and account for almost 9%. Finally, a few negations (2.58%) are multitoken (e.g., no longer, not at all), and several negations (almost 9%) appear in sentences with at least one more negation.

The corresponding affirmative interpretation is never just the original sentence with negation after removing the negation cue—doing so results in a sentence that is not semantically equivalent. The required modification are sometimes relatively simple and mainly require swapping a verb or adjective. For example, *are still not integrated* becomes *have yet to integrate*, *largely untouched* becomes *quite intact*, and *is useless* becomes *is futile*. Yet the affirmative interpretation often is a more thorough rewrite of the original sentence:

- *I have no doubt that we will reach our goal.* becomes *We shall surely get there!*;
- *higher precision that could never be achieved by hand* becomes *more precision than anyone performed manually; and*
- *lack of skilled labour* becomes *issue of obtaining sufficiently qualified personnel.*

4 Experiments with Natural Language Understanding Tasks

We leverage our collection of sentences with negation and their affirmative interpretations to enhance models for three natural language understanding tasks. First, we leverage them in a blending training setup to generate affirmative interpretations from the negations in a previous benchmark (Section 4.1). Second, we leverage them to create new premise-hypothesis pairs and build more robust models for natural language inference (Section 4.2). Third, we use them to train a plug-and-play neural component to generate affirmative interpretations from negation. Then, we incorporate the generator into the task of sentiment analysis (Section 4.3). We use existing corpora for all tasks as described below; for natural language inference and sentiment analysis we use the versions released by the GLUE benchmark (Wang et al., 2018).

Our experimental results show that leveraging the large collection of negations and their affirmative interpretations improves results across all tasks using previously proposed benchmarks. Specifically, we obtain either slightly better or comparable results when negation is not present in the input, and always better results when negation is present.

4.1 Generating Affirmative Interpretations from Negation

There are a couple corpora with sentences containing negation and their manually curated affirmative interpretations (Section 2). In our first experiment, we explore whether leveraging our collection is beneficial to generate affirmative interpretations from the negations in AFIN (Hossain et al., 2022b), a manually curated corpus that is publicly available. AFIN contains 3,001 sentences with negation and their affirmative interpretations. Unlike our collection (Section 3), AFIN only considers verbal negations (i.e., the negation cues always modify a verb). Here are some examples:

- *It was not formed by a natural process.*
  It was formed by an artificial process.
- *An extinct volcano is one that has not erupted in recent history.*
  An extinct volcano erupted in the past.

The AFIN authors experiment with the T5 (Raffel et al., 2020) transformer to automatically generate affirmative interpretations. While the task remains a challenge and our results are much worse than the human upper bound, we show that incorporating our collection of sentences containing negation and their affirmative interpretations during the training process results in a more robust generator.

Blending Our Collection of Negations and Affirmative Interpretations We adopt a blending technique by Shnarch et al. (2018) in order to maximize the chances that the training process benefits from our collection of negations and affirmative interpretations. Since our collection is much larger than the training split in AFIN (153k vs. 2.1k), simply adding our collection to the training split and fine-tuning T5 as usual would result in a model that underperforms with AFIN. There are three phases in the training process. In the first phase, we fine-tune T5 with the combination of our collection and

	BLEU	chrf++	METEOR
T5 transformer	26.5	50.5	43.5
+ blending Ours	28.6	52.5	45.8

Table 3: Automatic evaluation of the T5 transformer on the task of generating affirmative interpretations as defined in the AFIN benchmark (Hossain et al., 2022b). Blending our pairs in the process of fine-tuning T5 yields improvements with all metrics.

	Validation Scores				
	4	3	2	1	0
Upper Bound	86.2	11.6	2.0	0.2	n/a
T5 transformer	32.0	15.3	12.0	3.3	37.3
+ blending Ours	37.0	14.0	13.0	10.0	26.0

Table 4: Manual evaluation of the T5 transformer on the task of generating affirmative interpretations as defined in the AFIN benchmark. The upper bound comes from the manual validation by the creators of AFIN. Blending our pairs in the fine-tuning process generates more correct interpretations (higher validation scores are better).

the training split in AFIN for m epochs. In the second phase, we continue to fine-tune T5 blending our collection and the training split in AFIN for n epochs. The blending factor ($[0..1]$) determines the number of instances from our collection that we incorporate in each epoch. This number decreases after each epoch. In the third phase, we fine-tune using the training split in AFIN for k epochs. We refer the reader to Appendix B for additional details on the training process and hyperparameters.

Results and Discussion Table 3 presents the evaluation with the test split in AFIN using automatic metrics: BLEU-2 (Papineni et al., 2002), chrf++ (Popović, 2017), and METEOR (Banerjee and Lavie, 2005). We obtained these scores comparing the gold affirmative interpretations in AFIN and the predicted ones by T5. Despite our collection of negations and affirmative interpretation is noisy, out-of-domain, and considers more negation types, leveraging it is beneficial. Indeed, we observe improvements across the three metrics (BLEU-2: 28.6 vs. 26.5, chrf++: 52.5 vs. 50.5, and METEOR: 45.8 vs. 43.5).

Manual Validation Automatic metrics for generation tasks are useful but have well-known limitations (Mathur et al., 2020; Zhang et al., 2020). Following the AFIN authors, we also conduct a manual evaluation. Specifically, we validate a sample of 100 automatically generated affirmative interpretations. The validation consists in assigning a score indicating how confident they are in the correctness of an affirmative interpretation given the sentence containing negation (4: extremely confident, 3: very confident, 2: moderately confident, 1: slightly confident). We also include 0 to indicate that the affirmative interpretations is wrong. We show examples of each score in Appendix B.

Table 4 shows the manual evaluation. Blending our collection of negations and affirmative interpretations yields better results. While still far from the upper bound, blending increases the confidence scores. Most notably, the percentage of incorrect affirmative interpretations decreases from 37.3% to 26.0% ($\Delta = -30.3\%$).

4.2 Natural Language Inference

Our collection of pairs of sentences with negation and their affirmative interpretations can be seen as semantically equivalent sentences in which only one statement contains negation. By definition, there are two entailment relationships between semantically equivalent sentences—using either sentence as premise and the other one as hypothesis. We thus create two premise-hypothesis sets from each pair in our collection and label them as *entailment* to create a large collection of entailments involving negation. For example, we generate the following entailments from the pair (*The universal nature of these rights and freedoms does not admit doubts*, *The universal nature of these rights and freedoms is beyond question*):

- Premise: *The universal nature of these rights and freedoms does not admit doubts.*
  Hypothesis: *The universal nature of these rights and freedoms is beyond question.*
- Premise: *The universal nature of these rights and freedoms is beyond question.*
  Hypothesis: *The universal nature of these rights and freedoms does not admit doubts.*

This process results in 306,546 new premise-hypothesis annotated *entailment* (2 per pair in our collection) without any manual effort.

We experiment with (a) three transformer-based classifiers without any fine-tuning designed to improve results when there is a negation in the premise or hypothesis, (b) BERTNOT (Hosseini et al., 2021), a BERT transformer pretrained with a modified loss calculated in part with automatically

	RTE		SNLI		MNLI	
	dev	neg. P-H	dev	neg. P-H	dev	neg. P-H
w/o negation fine-tuning						
BERT (Devlin et al., 2019)	66.10	57.60	89.90	44.40	83.20	63.90
XLNet (Yang et al., 2019)	69.90	60.90	90.60	51.50	86.70	66.30
RoBERTa (Liu et al., 2019)	75.80	62.50	91.60	51.90	87.90	66.70
w/ negation fine-tuning						
BERTNOT (Hosseini et al., 2021)	69.68	74.47	89.00	45.96	84.31	60.89
RoBERTa blending Ours	77.62	78.13	91.35	54.87	87.00	67.89

Table 5: Results (accuracy) using several transformers and (a) the development splits in RTE, SNLI, and MNLI, and (b) the *new* premise-hypothesis containing negation (neg. P-H) from Hossain et al. (2020b). RoBERTa blending new premise-hypothesis derived from our sentences with negation and their affirmative interpretations substantially outperforms the three transformers without any negation fine-tuning and BERTNOT with the new premise-hypothesis that contain negation while obtaining comparable results with the original development splits.

obtained negated statements, and (c) a RoBERTa-based classifier blending our 306k new *entailment* premise-hypothesis using the strategy presented in Section 4.1. We refer the reader to Appendix C for additional details about the models, training process, and hyperparameters. Regarding corpora, we work with RTE (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). Additionally, we work with the 4,500 new premise-hypothesis pairs by Hossain et al. (2020b), who derive them from RTE, SNLI and MNLI by adding a negation to a premise, hypothesis or both.

Results and Discussion Table 5 presents the results. We train all models with the corresponding training split, except *RoBERTa blending Ours*, which also blends our 306k new premise-hypothesis pairs during the training process. We present results with the corresponding development split (gold labels for the test split are not available for all them) and the premise-hypothesis including negation. We find that blending our 306k premise-hypothesis is beneficial despite these pairs (a) only include entailments and (b) inherit the errors present in our collection of sentences with negation and their affirmative interpretations. With the original development splits in RTE, SNLI, and MNLI, we either obtain slightly better results (RTE, +1.82) or comparable (SNLI: -0.25, MNLI: -0.90). The improvements are consistent, however, with the pairs that include negation (neg. P-H). RTE and SNLI benefit the most (78.13 vs. 62.50, 54.87 vs. 51.90). We hypothesize that MNLI benefits the least (67.89 vs. 66.70) because premises in MNLI are often multiple sentences and our new premise-hypothesis are always single sentences.

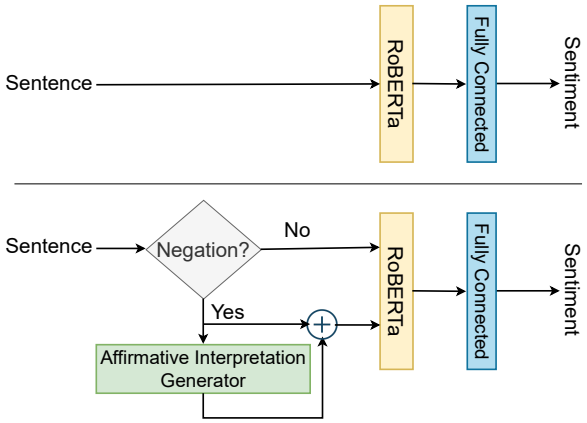

Figure 2: Standard architecture for a transformer-based classifier (top) and modification to include our affirmative interpretation generator (bottom). If a sentence contains a negation, we provide RoBERTa with the original sentence and the affirmative interpretation. The generator is pretrained with our large collection of sentences with negation and their affirmative interpretations.

4.3 Sentiment Analysis

We close our experiments exploring the task of sentiment analysis (i.e., classifying sentences according to their sentiment: positive or negative). Our motivation is that state-of-the-art systems for sentiment analysis face challenges with negation (Ribeiro et al., 2020). Unlike generating affirmative interpretations (Section 4.1) and natural language inference (Section 4.2), however, it is unclear how to leverage our large collection of sentences with negation and their affirmative interpretations to alleviate the issue.

We propose a task-agnostic solution: complement input sentences containing negation with their automatically generated affirmative interpretations. Figure 2 illustrates this proposal in the architecture at the bottom. Rather than feeding all sentences to a transformer-based classifier (top), we first check whether input sentences contain a negation with our cue detector (Section 3). If they do not, we feed them to the classifier as usual. If they do, we (a) automatically generate its affirmative interpretation with the generator described in Section 4.1 and (b) feed to the transformer-based classifier both the original sentence with negation and the affirmative interpretation separated by the [SEP] token. Appendix D provides additional details about the training process.

We experiment here with RoBERTa as it produces very competitive results. Note that our strategy to complement negated inputs with their affirmative interpretations could be used with any classifier for any task as long as it takes a text as its input. Regarding corpora, we use SST-2 (Socher et al., 2013) as released by GLUE (Wang et al., 2018). It consists of 70,042 movie reviews and sentiment annotations for each sentence.

Here are a few examples of automatically generated affirmative interpretations from sentences containing negation in SST-2:

- *It is not a bad film.*
  Affirmative interpretation: *It is a good movie.*
- *She may not be real, but the laughs are.*
  Affirmative interpretation: *She is fictional.*
- *He feels like a spectator and not a participant*
  Affirmative interpretation: *He feels like a spectator rather than participant.*
- *The movie has no idea of it is serious.*
  Affirmative interpretation: *The movie has a lack of idea that it is serious.*
- *A thriller without a lot of thrills.*
  Aff. interpretation: *A thriller with little thrills.*

Note that they are by no means perfect, but they mostly preserve meaning while not using negation. For example, the second affirmative interpretation only covers the meaning of part of the original sentence with negation (i.e., *She may not be real*), and the second to last includes a negation (*lack*).

Results and Discussion Table 6 presents the results (macro F1). We provide results with the sentences in the development split depending on whether they contain a negation (gold labels for the test split are not publicly available). Additionally, we use the grouping of negations by Hossain et al. (2022a): important or unimportant. A negation is unimportant if removing it does not change the sentiment (e.g., both *I got a headache watching this meaningless downer* and *I got a headache watching this downer* are negative.

	affirmative intpn. generator?	
	No	Yes
dev. w/o neg.	94.0	94.7 (+0.7%)
dev. w/ neg.	93.0	94.8 (+1.9%)
important negs.	86.0	89.8 (+4.4%)
unimportant negs.	95.0	95.8 (+0.8%)

Table 6: Results (macro F1) with RoBERTa using the SST-2 development split. We provide results with instances that contain and do not contain negation as well as important and unimportant negations. Our affirmative interpretation generator yields improvements across the board, especially with instances containing important negations (i.e., when removing the negation changes the sentiment polarity (positive or negative).

Incorporating our affirmative interpretation generator is always beneficial. This includes instances containing negation (94.8 vs. 93.0) and, surprisingly, instances that do not contain negation (94.7 vs. 94.0). We hypothesize that this is the case because our negation cue detector is not perfect thus instances are sometimes fed through the incorrect branch after the *Negation?* fork (Figure 2). As one would expect, the generator makes the classifier more robust with important negations (89.8 vs. 86), but we also observe improvements with unimportant negations (95.8 vs. 95.0).

5 Conclusions

Negation poses a challenge for natural language understanding. Understanding negation requires humans to infer affirmative meanings (Horn, 1989) (e.g., *The lot has not been vacant* conveys *The lot has been occupied*). Inspired by this insight, we collect a large collection (153k) of pairs of sentences containing negation and their affirmative interpretations. We define the latter as a semantically equivalent sentence that does not contain negation. Our collection process relies on parallel corpora and backtranslation and is automated.

We show that leveraging our collection is beneficial to solve three natural language understanding tasks: (a) generating affirmative interpretations, (b) natural language inference, and (c) sentiment

analysis. All our experiments use out-of-domain, manually curated corpora. Crucially, our proposal yields better results when negation is present in the input while slightly improving or obtaining comparable results when it is not. Additionally, our proposal does not require any manual annotations.

Limitations

In order to create a large collection of sentences with negation and their affirmative interpretations, we use publicly available parallel sentences. We note that in two of the three sources (i.e., WikiMatrix (Schwenk et al., 2021a) and CCMatrix (Schwenk et al., 2021b; Fan et al., 2021)), the authors use auto-alignment methodologies to collect the parallel sentences. This step may introduce errors in the original sources. Next, to detect negation cues in the huge collections of parallel sentences (14.6 millions), we develop a multilingual cue detection system that is certainly not 100% perfect. While the cue detector performs well on the negation corpora it is trained with (Section 3.2), some incorrect predictions can be expected in the parallel corpora we use. Furthermore, the translation API introduces additional noise backtranslating from Norwegian or Spanish into English (Section 3.3). Regarding models and experiments, we leverage RoBERTa and T5 (Section 4) as systems based on them perform well on natural language understanding tasks.[4] However, we acknowledge that other transformers such as XLNet (Yang et al., 2019) and DeBERTa (He et al., 2021) may yield better results.

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. Computational resources were provided by the UNT office of High-Performance Computing. Further, we leveraged computational resources from the Chameleon platform (Keahey et al., 2020). We also thank the reviewers for insightful comments.

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

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

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.

Eduardo Blanco and Dan Moldovan. 2011. Semantic representation of negation using focus detection. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–589, Portland, Oregon, 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*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Long Chen. 2019. Attention-based deep learning system for negation and assertion detection in clinical notes. *International Journal of Artificial Intelligence and Applications (IJAIA)*, 10(1).

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.

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.

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.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Federico Fancellu, Adam Lopez, and Bonnie Webber. 2016. Neural networks for negation scope detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 495–504, Berlin, Germany. Association for Computational Linguistics.

Federico Fancellu, Adam Lopez, Bonnie Webber, and Hangfeng He. 2017. Detecting negation scope is easy, except when it isn't. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 58–63, Valencia, Spain. Association for Computational Linguistics.

Margot Fonteyne, Arda Tezcan, and Lieve Macken. 2020. Literary machine translation under the magnifying glass: Assessing the quality of an NMT-translated detective novel on document level. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3790–3798, Marseille, France. European Language Resources Association.

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.

Uri Hasson and Sam Glucksberg. 2006. Does understanding negation entail affirmation?: An examination of negated metaphors. *Journal of Pragmatics*, 38(7):1015–1032.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Laurence Horn. 1989. *A Natural History of Negation*. University of Chicago Press.

Md Mosharaf Hossain, Dhivya Chinnappa, and Eduardo Blanco. 2022a. An analysis of negation in natural language understanding corpora. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 716–723, Dublin, Ireland. Association for Computational Linguistics.

Md Mosharaf Hossain, Kathleen Hamilton, Alexis Palmer, and Eduardo Blanco. 2020a. 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.

Md Mosharaf Hossain, Luke Holman, Anusha Kakileti, Tiffany Iris Kao, Nathan Raul Brito, Aaron Abraham Mathews, and Eduardo Blanco. 2022b. A question-answer driven approach to reveal affirmative interpretations from verbal negations. *arXiv preprint arXiv:2205.11467*.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020b. 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.

Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville. 2021. Understanding by understanding not: Modeling negation in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1301–1312, Online. Association for Computational Linguistics.

Rodney D. Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.

Liwei Jiang, Antoine Bosselut, Chandra Bhagavatula, and Yejin Choi. 2021. "I'm not mad": Commonsense implications of negation and contradiction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4380–4397, Online. Association for Computational Linguistics.

Salud María Jiménez-Zafra, Mariona Taulé, M Teresa Martín-Valdivia, L Alfonso Urena-López, and M Antónia Martí. 2018. Sfu reviewsp-neg: a spanish corpus annotated with negation for sentiment analysis. a typology of negation patterns. *Language Resources and Evaluation*, 52(2):533–569.

Salud María Jiménez-Zafra, Noa P. Cruz-Díaz, Maite Taboada, and María Teresa Martín-Valdivia. 2021. Negation detection for sentiment analysis: A case study in spanish. *Natural Language Engineering*, 27(2):225–248.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? on the ability of LSTMs to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231, Brussels, Belgium. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, 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.

Hao Li and Wei Lu. 2018. Learning with structured representations for negation scope extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 533–539, Melbourne, Australia. 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*.

Petter Mæhlum, Jeremy Barnes, Robin Kurtz, Lilja Øvrelid, and Erik Velldal. 2021. Negation in Norwegian: an annotated dataset. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 299–308, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Roser Morante and Walter Daelemans. 2012a. 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-2012)*. European Language Resources Association (ELRA).

Roser Morante and Walter Daelemans. 2012b. 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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.

Zahra Sarabi, Erin Killian, Eduardo Blanco, and Alexis Palmer. 2019. A corpus of negations and their underlying positive interpretations. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 158–167, Minneapolis, Minnesota. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Longxiang Shen, Bowei Zou, Yu Hong, Guodong Zhou, Qiaoming Zhu, and AiTi Aw. 2019. Negative focus detection via contextual attention mechanism. 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 2251–2261, Hong Kong, China. Association for Computational Linguistics.

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Melbourne, Australia. Association for Computational Linguistics.

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.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Hung Thinh Truong, Yulia Otmakhova, Timothy Baldwin, Trevor Cohn, Karin Verspoor, and Jey Han Lau. 2022. Not another negation benchmark: The nan-nli test suite for sub-clausal negation. *arXiv preprint arXiv:2210.03256*.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(11):S9.

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, 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, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yiyun Zhao and Steven Bethard. 2020. How does BERT's attention change when you fine-tune? an analysis methodology and a case study in negation scope. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4729–4747, Online. Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

Bowei Zou, Guodong Zhou, and Qiaoming Zhu. 2014. Negation focus identification with contextual discourse information. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 522–530, Baltimore, Maryland. Association for Computational Linguistics.

Bowei Zou, Guodong Zhou, and Qiaoming Zhu. 2015. Unsupervised negation focus identification with word-topic graph model. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1632–1636, Lisbon, Portugal. Association for Computational Linguistics.

A Additional Details on Identifying Negation Cues in Multiple Languages

Referring to Section 3.2 of the paper, we employ an off-the-shelf multilingual BERT-Base model (cased version) pretrained on 104 languages.[5] We concatenate the contextualized representations from the last and third-to-last layers. Then, we pass the concatenation to a fully connected layer. Finally, we leverage a conditional random field (CRF) layer that yields the output sequence identifying negation cues. Since a negation cue can consist of multiple tokens (e.g., *by no means*), we use the BIO (B: Beginning, I: Inside, and O: Outside)

[5] https://github.com/google-research/bert/blob/master/multilingual.md

Hyperparameter	
Maximum Epochs	25
Batch Size	10
Patience	5
Maximum sentence length	150
Optimizer	AdamW
Learning rate (mBERT)	1e-5
Learning rate (FC and CRF)	1e-3
Weight decay (mBERT)	1e-5
Weight decay (FC)	1e-3
Dropout (mBERT)	0.5
Gradient clipping	5.0
Warmup epochs	5
Accumulate step	1

Table 7: Hyperparameters for finetuning the multilingual cue detector (Section 3.2 in the paper). FC refers to Fully Connected layer.

Hyperparameter	
Maximum Epochs	20
Batch Size	8
Patience	5
Input sentence length (max.)	80
Target sentence length (max.)	50
Optimizer	Adafactor
Learning rate	1e-5
Weight decay	1e-6
Gradient clipping	5.0
Warmup epochs	3
Accumulate step	1
top_k	50
top_p	0.95
repetition_penalty	2.5

Table 8: Hyperparameters for finetuning our affirmative interpretation generator (Section 4.1 in the paper).

tagging scheme. The system takes 1.5 hours on average to train on a single core NVIDIA Tesla V100 (32GB). Table 7 lists the tuned hyperparameters for the cue detector. We avail the code for all our experiments at https://github.com/mosharafhossain/large-afin-and-nlu.

B Additional Details on Generating Affirmative Interpretations from Negation

We utilize the Huggingface implementation (Wolf et al., 2020) of T5, a conditional generation model (Section 4.1). In each run, the system requires approximately 7.2 hours to train on a single core NVIDIA Tesla V100 (32GB). Table 8 shows the hyperparameters for this experiment. Regarding Section 4.1 in the paper (Manual Validation), we present examples of each score in Table 9.

C Additional Details on Solving Natural Language Inference

We evaluate the systems on the development splits of the NLI benchmarks we work with (Section 4.2) as test split labels are not publicly available. So, we randomly select 15% examples of the original training split in order to tune the hyperparameters and to select the best model during the training process for each benchmark. We note that we evaluate on the development split with matched genres for MNLI. In each run, the system (blending with ours) requires approximately 2.1 hours to train for RTE, 7.8 hours for SNLI, and 9.6 hours for MNLI on a single core NVIDIA Tesla V100 (32GB). Table 10 presents the hyperparameters we use in this experiment.

D Additional Details on Solving Sentiment Analysis

To experiment with SST-2, we randomly select 5% examples of the original training split for tuning the hyperparameters as well as for selecting the best model during the training process since test labels are not publicly available in SST-2 (part of GLUE (Wang et al., 2018)). On average, the system takes half an hour to train on a single core NVIDIA Tesla V100 (32GB). We share the tuned hyperparameters in Table 11.

	Examples
Extremely confident (Score: 4)	Move them to low places so that they do not fall. Affirmative Interpretation: They fall when they are in higher places. The ones that were not rewarded were not marked with fields. Affirmative Interpretation: The ones that were rewarded were marked with fields.
Very confident (Score: 3)	The most recent successful bids for the Olympic and Paralympic Games were in cities that had never hosted them before. Affirmative Interpretation: Other cities had hosted them once. No other studies could find a link between the vaccine and autism. Affirmative Interpretation: A study found a link between the vaccine and autism.
Moderately confident (Score: 2)	In 1984, because the games were in New York, and because of the boycott, from when we boycotted in 1980, not a lot of European countries came over. Affirmative Interpretation: Lots of European countries came over in 1980. It occurs when the body does not receive enough iron. Affirmative Interpretation: The body receives too little iron.
Slightly confident (Score: 1)	I understand third party candidates have no success. Affirmative Interpretation: Third party candidates have minimal success. I don't expect that the lack of British participation will stop any action. Affirmative Interpretation: I expect that the lack of British participation will slow down any action.
Wrong affirmative interpretation (Score: 0)	I have throughout my career not supported needle exchanges as anti-drug policies. I have supported needle exchanges as anti-drug policies. Unlike other organelles, the ribosome is not surrounded by a membrane. The ribosome is surrounded by a membrane.

Table 9: Examples of sentences containing negation from AFIN and their affirmative interpretations automatically generated. The generator uses T5 trained with our large collection of sentences with negation and their affirmative interpretations (Section 4.1). We show examples of the manual validation; scores range from 0 to 4.

Hyperparameter	RTE	SNLI	MNLI
Maximum epochs	10	8	8
Warmup epochs	4	2	3
Batch size	24	16	10
Patience	5	5	5
Maximum sentence length	100	80	100
Optimizer	AdamW	AdamW	AdamW
Learning rate	1e-5	1e-5	1e-5
Weight decay	0.0	5e-6	5e-6
Gradient clipping	5.0	5.0	5.0
Dropout	0.2	0.3	0.3

Table 10: Hyperparameters for finetuning RoBERTa with blending our pairs and the NLI benchmarks (Section 4.2 in the paper).

Hyperparameter	
Maximum Epochs	5
Batch Size	16
Patience	3
Maximum sentence length	80
Optimizer	AdamW
Learning rate	1e-5
Weight decay	0.0
Dropout	0.5
Gradient clipping	5.0
Warmup epochs	5
Accumulate step	1

Table 11: Hyperparameters for finetuning our SST-2 system presented in Section 4.3 in the paper.