Improving negation detection with negation-focused pre-training

Hung Thinh Truong[1] **Timothy Baldwin**[1,3] **Trevor Cohn**[1] **Karin Verspoor**[2]
[1]The University of Melbourne, [2]RMIT University, [3]MBZUAI
hungthinht@student.unimelb.edu.au, tb@ldwin.net,
trevor.cohn@unimelb.edu.au, karin.verspoor@rmit.edu.au

Abstract

Negation is a common linguistic feature that is crucial in many language understanding tasks, yet it remains a hard problem due to diversity in its expression in different types of text. Recent work has shown that state-of-the-art NLP models underperform on samples containing negation in various tasks, and that negation detection models do not transfer well across domains. We propose a new negation-focused pre-training strategy, involving targeted data augmentation and negation masking, to better incorporate negation information into language models. Extensive experiments on common benchmarks show that our proposed approach improves negation detection performance and generalizability over the strong baseline NegBERT (Khandelwal and Sawant, 2020).

1 Introduction

Negation is an important linguistic phenomenon that appears commonly in natural language but is underrepresented in common NLP benchmarks (Hossain et al., 2020). Furthermore, the Checklist benchmark (Ribeiro et al., 2020) shows that most sentiment analyzers and machine comprehension models struggle with samples containing negation. Negation is even more important in biomedical domain text, where patients are carefully defined as having/not having specific characteristics. Even within the biomedical domain, there are many types of text such as clinical notes, lab reports, or research publications, each with particular characteristics in relation to the use of negation. A recent study on English texts found that negation detection models do not transfer well across domains, due to variations in expression of negation (Khandelwal and Sawant, 2020). It remains a challenge to solve negation in general, even with state-of-the-art NLP models.

Negation detection is typically defined as consisting of the two sub-tasks of: (1) cue detection, detecting the cue phrase that triggers the negation; and (2) scope resolution, determining the affected spans that are negated. There are three primary datasets that have been used to evaluate negation: (1) the *BioScope* corpus (Vincze et al., 2008) includes full papers and abstracts of biological papers; (2) the *SFU* corpus (Konstantinova et al., 2012) is a collection of product reviews; and (3) the *Sherlock* dataset (Morante and Blanco, 2012) consists of short literary works. There are differences in annotation schemes across the datasets, such as whether or not the cues are included inside scope annotation, and sub-optimal cross-dataset results have been observed, providing clear indications that the datasets are highly divergent in language use and negation types.

In this work, we aim to extend the transfer learning capability of NegBERT (Khandelwal and Sawant, 2020) through additional pre-training with task-related augmented training data, and a new masking objective. Our contributions are:

- We introduce an approach to augmenting data to emphasize negation in pre-training.
- We propose a novel extension to the standard random masked language model objective in pre-training to explicitly mask negation cues, to make the models more robust to negation.
- We conduct extensive experiments on different benchmarks to evaluate cross-domain performance of large pre-trained language models as well as the effectiveness of the proposed pre-training strategies; code is available at https://github.com/joey234/negation-focused-pretraining.

2 Related work

To date, negation detection has been heavily reliant on rule-based systems. Chapman et al. (2001) proposed a simple system, NegEx, based on regular expressions to detect negation cues in a sentence given a concept of interest (the scope). NegEx re-

mains the most popular approach to negation detection, especially in the clinical domain to determine the polarity of clinical concepts (e.g., as sourced from MetaMap (Aronson and Lang, 2010)). Further research has extended NegEx with syntactic information (Mehrabi et al., 2015; Peng et al., 2018), and shown that rule-based systems can achieve relatively good performance for detecting negation, especially in the biomedical domain, but do not generalize well to other domains or datasets.

To approach negation cue and negation scope detection with supervised machine learning, two classification tasks are defined: (1) finding negation tokens, and (2) classifying tokens as the first or last (or neither) token within the scope of negation. Most work follows a common scheme in extracting various features from the sentence, and using a classifier to classify each token as the beginning, inside, or outside of a negation cue or scope span (Morante and Daelemans, 2009; Ou and Patrick, 2015; Cruz et al., 2016). Recently, research has shifted to applying deep learning methods to the task. Most approaches make use of RNN-based architectures to encode the input sentences, combined with a softmax layer for classification (Lazib et al., 2019; Chen, 2019). Despite the high performance on common benchmarks, results are biased by the fact that negation scope is often delimited by punctuation and other dataset artefacts (Fancellu et al., 2017). As such, they are potentially only learning domain-specific surface features rather than capturing the true semantics of negation. NegBERT applies a large pre-trained language model to the problem of negation detection, outperforming previous deep learning methods on negation detection, with especially high gains on scope resolution.

3 Method

Our proposed pre-training strategy consists of two main components: (1) negation-focused data collection in which we first collect relevant data that contains negation; and (2) negation-focused pre-training that makes use of the negation-focused data to emphasize negation instances, and adopts a novel negation-specific masking strategy.

3.1 Negation-focused data collection

We aim to construct a dataset that is enriched with negation information, to support negation-sensitized pre-training of large language models. To obtain sentences with negations, we extend the NegEx lexicon with additional negation cues obtained from biomedical texts (Morante, 2010), and apply it to sentences extracted from a corpus using the SpaCy English sentence tokenizer, keeping only those sentences with at least one identified negation cue.

For the biomedical domain, we use texts in the TREC-CDS 2021 snapshot[1] of the clinical trials registry.[2] Clinical trials are documents describing the protocols and relevant patient characteristics of a clinical research study. Descriptions of clinical trials can be quite long, but a core aspect of the trial description is the patient inclusion/exclusion criteria, specifying what types of characteristics or conditions a patient must have/not have in order to be suitable for the trial. The reasons for choosing this data are that: (1) it is in-domain for the biomedical domain; (2) the texts are well-formed sentences with proper grammatical structure; and (3) the texts contain many negations, especially in the inclusion/exclusion criteria sections. For the general domain, we apply this approach to *wikitext* (Merity et al., 2016), a set of verified articles in Wikipedia. We sample the data equally from these two sets, obtaining $1,381,948$ negation sentences.

3.2 Negation-focused pre-training

Adaptive pre-training on target domain data has been shown to be an effective strategy for domain adaptation (Gururangan et al., 2020). We therefore hypothesize that pre-training language models on text with negations will help the model incorporate information about negation, and learn better representations for sentences containing negation. Using the negation-focused data, we first apply the standard random word masking strategy (Devlin et al., 2019) and train the model with the masked language model objective.

As part of the collection of the negation-focused data, we obtain predictions of negation cues in all the sentences, which can be explicitly incorporated to make the model more robust to negation. Inspired by work on entity and span masking (Joshi et al., 2020; Yamada et al., 2020), we explore explicitly incorporating information about negation cues into the model by masking these cues, and targeting prediction of the masked cue in the pre-training stage. Below is an example of how a sentence is tokenized under our masking scheme:

[1] http://www.trec-cds.org/2021.html
[2] http://clinicalTrials.gov

No serious complications such as hypertension, diabetes. ⇒ ***[CUE]*** *serious complications such as* ***[MASK]****, diabetes.*

A new type of token ***[CUE]*** is introduced under this masking scheme, and the model needs to reconstruct the original sentence by predicting both the ***[CUE]*** token to be *No*, and the randomly-masked token ***[MASK]*** to be *hypertension*. By always masking negation cues in all the sentences, we force the model to focus more on this type of token, and thus, aim to learn better embeddings incorporating information of how a negation cue is represented in the context of the sentence. Moreover, by using a different token to mask negation cues, we ensure that the model learns to distinguish between different types of tokens. In this work, we replace the BERT encoder of NegBERT with RoBERTa (Liu et al., 2019) and apply whole-word masking, meaning that all the sub-word tokens that constitute a word will be masked.

4 Experiments

4.1 Experimental settings

Following the experimental settings in NegBERT, we use the three standard benchmarks for negation cue detection and scope resolution tasks, i.e. BioScope (Vincze et al., 2008) (separated into two subsets, sourced from abstracts and full-text papers, resp.), the SFU product reviews dataset (Konstantinova et al., 2012), and the Sherlock dataset (Morante and Blanco, 2012). In addition, we use the negation-annotated subset of VetCompass UK[3] (Cheng et al., 2017), consisting of clinical notes in the veterinary domain, which are very informal compared to BioScope. It also contains abbreviations and shortening of terms, as well as certain unique negation cues. To investigate cross-domain performance, we perform cue detection and scope resolution for all 4 datasets, based on training on one dataset and evaluating on all datasets. Detailed statistics of these datasets are presented in Table 1. Note that we do not experiment with the setting of training with all the combined training data from the corpora as it has been pointed out by previous work that doing so hurts the performance of the models (Jiménez-Zafra et al., 2020; Barnes et al., 2021) due to differences in annotation schemes between the corpora introducing noise during training. Re-annotating all the datasets using a common annotation scheme would be a potential solution here, which we leave for future work.

[3] https://www.rvc.ac.uk/VetCOMPASS

Dataset	#sentences	#negations	#unique cues
BioScope-Abstract	11871	1719	28
BioScope-FullPaper	2670	376	18
SFU	17263	3527	53
Sherlock	5520	1421	30
VetCompass	6582	724	26

Table 1: Dataset statistics

We formulate the two tasks as sequence labeling problems, where each token is tagged with a corresponding label. For cue detection, we use the annotation scheme *{0: Affix, 1: Normal Cue, 2: Part of multiword cue, 3: Not part of cue}*. For scope resolution, we use gold cue information and two labels *{0: Outside negation scope, 1: Part of negation scope}*. We adopt the same hyperparameters as NegBERT. Following the standard evaluation scheme in previous negation detection works, all systems are evaluated using token-level F_1-score, based on whether it is inside or outside of any negation cue or scope. Methods evaluated include: (1) **NegBERT**; (2) **AugNB** = NegBERT plus pre-training on negation-focused data; and (3) **CueNB** = NegBERT plus pre-training on negation-focused data and the negation cue masking objective. Note that for NegBERT, we also replace the BERT encoder with RoBERTa to ensure results are comparable between the models.

4.2 Main results

Tables 2 and 3 report the performance of negation cue detection and negation scope resolution, respectively. Results reported are the average of 5 runs with different random seeds. NegBERT results are produced using the official implementation.[4] To provide a more general view, we summarize the results in Table 4. In general, we observe gains in both the same-dataset setting (training and test set belongs to one corpus) and cross-dataset setting (training one one training set and testing on all others test sets) for both of the proposed models, with CueNB achieving the largest gains.

We observe similar trends across all datasets for

[4] https://github.com/adityak6798/Transformers-For-Negation-and-Speculation

Evaluation set \ Training set	BioScope-Abstract			BioScope-FullPaper			SFU			Sherlock			VetCompass		
	NegBERT	AugNB	CueNB	NegBERT	AugNB	CueNB	NegBERT	AugNB	CueNB	NegBERT	AugNB	CueNB	NegBERT	AugNB	CueNB
BioScope-Abstract	95.29	+0.78	+1.80	92.58	−0.27	−0.25	83.86	+0.23	+0.47	70.67	+4.89	+6.62	75.49	+1.56	+2.73
BioScope-FullPaper	91.44	+0.68	+0.89	90.23	+0.04	+1.44	79.68	+0.11	+0.63	66.45	+1.07	+2.48	71.06	+1.17	+2.98
SFU	38.70	+3.55	+4.78	57.99	+3.72	+4.43	87.20	+0.38	+0.79	44.03	+11.37	+13.93	58.66	+0.67	+2.44
Sherlock	70.43	+2.86	+3.05	69.63	+4.54	+6.48	70.14	+1.26	+1.79	92.28	−0.51	−1.11	64.45	+3.58	+3.99
VetCompass	70.58	+0.37	+1.91	69.75	+0.36	+2.39	75.18	+2.19	+3.42	71.34	+0.33	+1.07	87.77	+1.11	+3.77

Table 2: Cue detection results. Gray cells denote the same-dataset setting, and green cells indicate the highest score for each evaluation dataset. Results of AugNB and CueNB are relative changes compared to NegBERT

Evaluation set \ Training set	BioScope-Abstract			BioScope-FullPaper			SFU			Sherlock			VetCompass		
	NegBERT	AugNB	CueNB	NegBERT	AugNB	CueNB	NegBERT	AugNB	CueNB	NegBERT	AugNB	CueNB	NegBERT	AugNB	CueNB
BioScope-Abstract	94.23	+0.84	+1.58	90.89	+0.70	+0.74	84.41	+0.15	+0.43	78.80	+0.63	+1.66	69.14	+1.57	+2.82
BioScope-FullPaper	91.63	+1.14	+1.83	88.42	+1.80	+4.14	79.90	+0.49	+0.83	79.42	+0.21	+1.28	64.45	+1.79	+2.27
SFU	85.28	+0.78	+1.03	84.57	+0.71	+1.05	90.44	+0.27	+0.59	74.61	+1.88	+3.28	63.32	+3.25	+3.59
Sherlock	72.60	+0.43	+2.17	70.10	+2.24	+3.04	73.68	+0.11	+0.87	91.51	−1.20	−0.27	61.49	−0.03	+1.64
VetCompass	61.36	+0.86	+2.00	60.27	+1.06	+1.39	62.62	+0.32	+1.40	59.62	+0.61	+1.05	88.18	+1.23	+2.06

Table 3: Scope resolution results. Gray calls denote the same-dataset setting, and green cells indicate the highest score for each evaluation dataset. Results of AugNB and CueNB are relative changes compared to NegBERT

Task	Same-dataset results			Cross-dataset results		
	NegBERT	AugNB	CueNB	NegBERT	AugNB	CueNB
Cue Detection	90.55	+0.36	+1.34	69.61	+2.21	+ 3.31
Scope Resolution	90.56	+ 0.59	+1.62	73.41	+0.95	+ 1.72

Table 4: Aggregated results

both cue detection and scope resolution. Regarding the in-dataset setting, AugNB outperforms the baseline NegBERT on all datasets except for Sherlock. Gains are more noticeable over the biomedical datasets (BioScope, VetCompass). For Sherlock, however, we observe a slight degradation in performance with the proposed pre-training scheme. This is likely due to the fact that Sherlock has major differences in annotation scheme compared to other corpora, specifically including scopes to the left of cues, while in BioScope and SFU, the scope is usually annotated only to the right of cues. Also, the cue itself is not considered to be part of the scope in Sherlock or SFU, unlike in BioScope.

In the cross-dataset setting, we record gains across all benchmarks. The largest cross-dataset improvements over NegBERT are for SFU, perhaps due to SFU being the largest dataset in size, containing a relatively large number of unique cues. CueNB further improves the performance of AugNB, confirming our hypothesis that explicitly masking the cue helps the model learn better representations for negation cues and thus, better distinguish between cues and normal words. These results show that our negation-focused pre-training strategy is effective for improving the transfer learning performance of pre-trained language models on the negation detection task.

4.3 Discussion

We conducted an error analysis on the VetCompass validation set to see what qualitative improvement CueNB makes over NegBERT. For cue detection, there are two main types of errors that CueNB helps alleviate. First, CueNB can detect more unique cues such as *negative*, *won't*, and also multiword cues like *no longer*. Second, CueNB is able to recognize cases when the negations are actually just speculative. For example, in the sentence *O reports has smelled for past week,* ***not sure*** *if anal glands* . . . , the word *not* is part of the speculation phrase *not sure*, indicating that this is not truly a negation phrase but rather expresses uncertainty. For scope resolution, CueNB mostly helps in recognizing the correct scope boundary. One common case is when the cue relates to multiple spans in a sentence. In the sentence *Examination: QAR, thorac ausc and abdo palp NAD*,[5] NegBERT only recognizes the nearest span *abdo palp NAD* to be the scope, whereas CueNB recognizes the full correct span *thorac ausc and abdo palp NAD*. It also helps in cases where there are multiple separate negations in the same sentence. For instance, in the

[5] *NAD* is the negation cue *no abnormality detected*, *QAR*, *thorac ausc* (thoracic ausculation), *abdo palp* (abdonimal palpation), are different types of physical examinations.

Model	Cue	Scope
NegBERT	94.46	95.34
+ negation-focused data	95.36	95.94
+ explicit cue masking	95.58	96.03
CueNB	**95.87**	**96.76**

Table 5: Ablation study on BioScope validation set.

sentence *No V+ or no D+.*, the phrases *No V+* and *no D+* are two independent negation scope spans, while NegBERT would recognize the whole sentence as a single span. Another interesting case is when there are exceptions in the sentence, e.g. the *No ... other than ...* construction. For *No probs detected other than the skin lesions*, CueNB is able to recognize the correct scope *No probs detected* while NegBERT considers the whole sentence to be the scope.

We also conduct an ablation study to understand the impact of each component of the proposed pre-training strategy. Table 5 presents the results of different variations of the proposed pre-training scheme on the BioScope-Abstract validation split. We consider two variations, pre-training with: (1) only the negation-focused data (equivalent to the AugNB model); and (2) only the cue masking objective. To model the latter variation, we explicitly mask the cue in the BioScope training set, then pre-train on this training set. From the results, we see that both strategies help improve the baseline NegBERT on cue detection and scope resolution, with explicitly masking the cues being the most important. Combining both strategies (CueNB) further improves the overall results.

5 Conclusion

In this work, we propose a new negation-focused pre-training strategy to explicitly incorporate negation information into pre-trained language models. Empirical results on common benchmarks show that the proposed strategy helps improve the performance of pre-trained language models on the negation detection task when evaluating on the same source dataset, as well as their transferability to target data in different domains. Despite the gains over previous methods, the sub-optimal results on some benchmarks show that negation remains a big challenge in NLP.

Acknowledgements

The authors would like to thank the anonymous reviewers for their constructive reviews. This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200. This research was conducted by the Australian Research Council Training Centre in Cognitive Computing for Medical Technologies (project number ICI70200030) and funded by the Australian Government.

References

Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Jeremy Barnes, Erik Velldal, and Lilja Øvrelid. 2021. Improving sentiment analysis with multi-task learning of negation. *Natural Language Engineering*, 27(2):249–269.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

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).

Katherine Cheng, Timothy Baldwin, and Karin Verspoor. 2017. Automatic negation and speculation detection in veterinary clinical text. In *Proceedings of the Australasian Language Technology Association Workshop 2017*, pages 70–78.

Noa P Cruz, Maite Taboada, and Ruslan Mitkov. 2016. A machine-learning approach to negation and speculation detection for sentiment analysis. *Journal of the Association for Information Science and Technology*, 67(9):2118–2136.

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, USA.

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.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

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.

Salud María Jiménez-Zafra, Roser Morante, María Teresa Martín-Valdivia, and L. Alfonso Ureña-López. 2020. Corpora annotated with negation: An overview. *Computational Linguistics*, 46(1):1–52.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

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.

Natalia Konstantinova, Sheila C.M. de Sousa, Noa P. Cruz, Manuel J. Maña, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3190–3195, Istanbul, Turkey. European Language Resources Association (ELRA).

Lydia Lazib, Yanyan Zhao, Bing Qin, and Ting Liu. 2019. Negation scope detection with recurrent neural networks models in review texts. *International Journal of High Performance Computing and Networking*, 13(2):211–221.

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*.

Saeed Mehrabi, Anand Krishnan, Sunghwan Sohn, Alexandra M Roch, Heidi Schmidt, Joe Kesterson, Chris Beesley, Paul Dexter, C Max Schmidt, Hongfang Liu, et al. 2015. DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. *Journal of Biomedical Informatics*, 54:213–219.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Roser Morante. 2010. Descriptive analysis of negation cues in biomedical texts. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

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

Roser Morante and Walter Daelemans. 2009. A metalearning approach to processing the scope of negation. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 21–29.

Ying Ou and Jon Patrick. 2015. Automatic negation detection in narrative pathology reports. *Artificial Intelligence in Medicine*, 64(1):41–50.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. 2018. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188.

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.

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):1–9.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454.