Negation Scope Conversion:
Towards a Unified Negation-Annotated Dataset

Asahi Yoshida[1], Yoshihide Kato[2], Shigeki Matsubara[1,2]
[1]Graduate School of Informatics, Nagoya University
[2]Information and Communications, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
yoshida.asahi.y7@s.mail.nagoya-u.ac.jp

Abstract

Negation scope resolution is the task that identifies the part of a sentence affected by the negation cue. The three major corpora used for this task, the BioScope corpus, the SFU review corpus and the Sherlock dataset, have different annotation schemes for negation scope. Due to the different annotations, the negation scope resolution models based on pre-trained language models (PLMs) perform worse when fine-tuned on the simply combined dataset consisting of the three corpora. To address this issue, we propose a method for automatically converting the scopes of BioScope and SFU to those of Sherlock and merge them into a unified dataset. To verify the effectiveness of the proposed method, we conducted experiments using the unified dataset for fine-tuning PLM-based models. The experimental results demonstrate that the performances of the models increase when fine-tuned on the unified dataset unlike the simply combined one. In the token-level metric, the model fine-tuned on the unified dataset archived the state-of-the-art performance on the Sherlock dataset.

1. Introduction

Negation is a common linguistic phenomenon in natural language that reverses the meaning of a sentence, phrase, word, etc. As Jiménez-Zafra et al. (2020) stated, negation has a remarkable impact on NLP systems. Therefore, detecting negations accurately is crucial for such systems. Many previous studies have addressed the automatic negation detection (Wu and Sun, 2023; Truong et al., 2022; Khandelwal and Sawant, 2020) and its application to downstream tasks (Moore and Barnes, 2021; Barnes et al., 2021; Mukherjee et al., 2021). Typically, the negation detection consists of two subtasks: (i) negation cue detection, detecting words that express negations (e.g., not, no) and (ii) negation scope resolution, resolving the part(s) of a sentence affected by the cue.

This study addresses negation scope resolution, which is still a hard problem due to its complexity. There are three primary corpora used for this task: the BioScope corpus (Szarvas et al., 2008), the SFU review corpus (Konstantinova et al., 2012) and the Sherlock dataset (Morante and Daelemans, 2012). When performing negation scope resolution with the model trained from the simply combined corpora, one issue arises: each corpus adopts different annotation schemes. Table 1 shows different annotations of negation cue and scope for the sentence "She does not have a cat.". The difference causes a decrease in the performance of negation scope resolution models. In fact, Barnes et al. (2021) reported the experiment

Corpus	Annotation of negation
BioScope	She does **not** <u>have a cat</u>.
SFU	She does **not** <u>have a cat</u>.
Sherlock	<u>She does</u> **not** <u>have a cat</u>.

Table 1: Annotations of negation cue (marked in bold) and scope (underlined) for the sentence "She does not have a cat." in the three corpora.

where simply combined dataset led to lower model performance. Truong et al. (2022) conjectured that merging different corpora into a unified dataset based on a common annotation scheme can be a potential solution for this issue. One possible approach for creating a unified dataset is to manually re-annotate the three corpora according to a common annotation scheme. However, such manual annotation requires expert linguists and consumes much time and effort. Automated conversion of negation scopes can be an alternative solution, but to the best of our knowledge, there is no previous study that has developed such a method.

Inspired by the points mentioned above, this paper proposes a method for automatically converting the scopes of BioScope and SFU to those of Sherlock[1]. We select the scope annotation of Sherlock as the target for conversion because it can represent more complex negation scopes compared to those of BioScope and SFU. Using the proposed method, we can obtain the variants

---

[1]Our code is available at https://github.com/asahi-y/negation-scope-conversion.

of BioScope and SFU whose annotation is almost identical to Sherlock. By merging Sherlock and the converted version of BioScope and SFU into a unified dataset, we have a large training data for negation scope resolution models. To verify the effectiveness of the proposed method, we conducted experiments of negation scope resolution using the unified dataset for fine-tuning the models. The experimental results show that the dataset created by our method improves the performance of the scope resolution model unlike the simply combined one. In particular, the model fine-tuned on the unified dataset archived the state-of-the-art performance in the token-level metric on the Sherlock dataset.

2. Negation-Annotated Corpora

This section describes the three negation-annotated corpora, which we use in this study.

BioScope The BioScope corpus (Szarvas et al., 2008)[2] contains biological text where negation cues are annotated with their scopes. Sentences (1) and (2) are examples annotated with negation cue (marked in bold) and its scope (underlined).

(1) The transcription factors did **not** change.

(2) The feature was **not** seen in the resting cells.

Basically, the scope annotation covers only the right of the cue. There are exceptional cases where the scope annotation also covers the left of the cue (e.g., a passive sentence is such a case as shown in sentence (2)).

SFU The SFU review corpus (Konstantinova et al., 2012)[3] is a collection of reviews. Most annotation schemes of negation scope follow those of BioScope.

Sherlock The Sherlock dataset (Morante and Daelemans, 2012)[4] contains Conan Doyle stories annotated with negation cues and their scopes. Sentences (3) and (4) are annotation examples.

(3) We did **not** drive up to the door.

(4) You'll see how **im**possible for me to go there.

The annotation guidelines of Sherlock (Morante et al., 2011) are based on those of BioScope, but there are several improvements. The main improvements are as follows:

- The scope includes all the arguments of the event being negated (e.g., the scope in sentence (3) includes the subject "we").

- Affixal negation cues are also considered (e.g., in sentence (4), the affixal cue "im" and part of its scope "possible" are distinguished).

- Discontinuous scopes are allowed.

These improvements make it possible to represent more complex negation scopes. Fancellu et al. (2017) pointed out that the scope annotation of Sherlock is linguistically motivated. Considering these points, this study explores a method for converting the scopes of BioScope and SFU to those of Sherlock.

3. Negation Scope Conversion

This section proposes a method for automatically converting the negation scopes of BioScope[5] and SFU to those of Sherlock. We use a negation scope resolution method as a basis of conversion. Below, the combination of BioScope and SFU is called B&S[6].

To utilize the correct scope annotations in B&S, our method converts negation scopes of B&S as follows:

$$S_{\text{left}} \cup S_{\text{cue}} \cup S_{\text{mid}} \cup S_{\text{right}}$$

where

$$S_{\text{left}} = \begin{cases} \mathsf{L}_c(S_{\text{B\&S}}) & (\mathsf{L}_c(S_{\text{B\&S}}) \neq \emptyset) \\ \mathsf{L}_c(S_{\text{res}}) & (\mathsf{L}_c(S_{\text{B\&S}}) = \emptyset), \end{cases}$$
$$S_{\text{mid}} = \mathsf{M}_c(S_{\text{B\&S}}), \; S_{\text{right}} = \mathsf{R}_c(S_{\text{B\&S}}).$$

Here, $S_{\text{B\&S}}$ and S_{res} represent the scopes of B&S and the result of a scope resolution method, respectively. c is a negation cue. S_{cue} represents internal structure of the cue and is defined in the next section. $\mathsf{L}_c(S)$, $\mathsf{M}_c(S)$ and $\mathsf{R}_c(S)$ are the left, middle[7] and right parts of the scope, respectively

---

[2] https://rgai.inf.u-szeged.hu/node/105
[3] https://www.sfu.ca/~mtaboada/SFU_Review_Corpus.html
[4] https://www.clips.ua.ac.be/sem2012-st-neg/data.html

[5] As in the previous studies, we use the two sub-corpora of BioScope: Abstract and FullPaper, because the original texts of the other sub-corpus (Clinical) is not publicly available.

[6] Unlike SFU and Sherlock, the scope annotation of BioScope includes the cue. To handle this format difference, we remove cues from scopes of BioScope as a pre-processing.

[7] $\mathsf{M}_c(S)$ exists only when the negation cue consists of multiple words and the words are discontinuous (e.g., **neither**, **nor**...).

*Surprisingly, the restaurant **didn't** <u>accept credit cards</u>.*

processing for the contracted cue (Sec. 3.1)	⬇

*Surprisingly, the restaurant did**n't** <u>accept credit cards</u>.*

scope resolution for the left part of the cue (Sec. 3.2)	⬇

*Surprisingly, <u>the restaurant did</u>**n't** <u>accept credit cards</u>.*
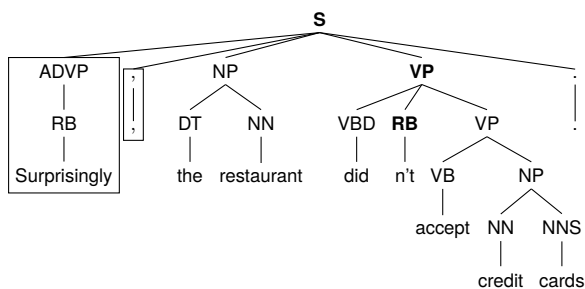
Figure 1: Example of scope conversion.

Figure 2: Parse tree of the sentence "Surprisingly, the restaurant didn't accept credit cards.", highlighting candidate constituents. Enclosed parts are removed in the adjustment step.

and are defined as follows:

$$L_c(S) = \{i \in S \mid i < c_l\},$$
$$M_c(S) = \{i \in S \mid c_l < i < c_r\},$$
$$R_c(S) = \{i \in S \mid c_r < i\}.$$

Here, c_l and c_r represent the positions in the sentence for the leftmost and rightmost words of the negation cues, respectively. The calculation is based on the following observations in the training data and the annotation guidelines of the three corpora:

1. $R_c(S_{\text{B\&S}})$ and $R_c(S_{\text{SH}})$, as well as $M_c(S_{\text{B\&S}})$ and $M_c(S_{\text{SH}})$ can be regarded as almost identical. Here, S_{SH} is the scope of the Sherlock dataset.

2. $L_c(S_{\text{B\&S}})$ and $L_c(S_{\text{SH}})$ can be regarded as almost identical if $L_c(S_{\text{B\&S}}) \neq \emptyset$ (e.g., sentence (2) is the case).

3.1. Affixal and Contracted Cues

Sherlock distinguishes the cue and its scope in the word for affixal cues (e.g., **un**<u>usual</u>) and contracted cues (e.g., <u>do</u>**n't**), whereas B&S treats whole the word as the cue. We distinguish affixal and contracted cues in B&S based on a simple pattern matching. If B&S's cue c has an affix in V_{aff}, or a suffix in V_{cont}, S_{cue} is a singleton set consisting of the result by removing the affix from the cue.

RB//VP/S/SBAR if SBAR\WHNP	UH
RB//VP/S	IN/PP
RB//S	<u>NN//NP/NP if lemma of NN in ["*lack*", "*absence*"]</u>
DT/NP if NP/PP	NN//NP//S/SBAR if SBAR\WHNP
DT//SBAR if SBAR\WHADVP	NN/NP//S
DT//S	CC/SINV
JJ//ADJP/VP/S if S\VP\VB* [@lemma="*be*"]	<u>VBG//NP</u>
JJ/NP/NP if NP\PP	<u>VBN//NP</u>
JJ//NP	<u>VB*//S</u>

X/Y : Y is the parent of X
X//Y : Y is an ancestor of X
X\Y : Y is a child of X

Figure 3: Path pattern rules. Each row represents one rule, which is shown in the order they are applied. The underlined rules are the ones we add in this study.

Based on the annotation guidelines of Sherlock (Morante et al., 2011) and the training data, we define V_{aff} and V_{cont} as follows:

$$V_{\text{aff}} = \{\text{dis}, \text{im}, \text{in}, \text{ir}, \text{un}, \text{less}\} \ , \ V_{\text{cont}} = \{\text{n't}, \text{not}\} \ .$$

3.2. Negation Scope Resolution

We obtain S_{res} using the scope resolution method of Yoshida et al. (2023) with several modifications, which adapt to the domains of BioScope and SFU. Below, we explain our modifications using an example shown in Figure 1 with its parse tree shown in Figure 2. The method follows three steps below. We modify steps 2 and 3.

1. Parse the sentence and select the constituents that dominate the cue as candidates.

2. From the candidates, select one constituent corresponding to the scopes using the heuristics.

3. Adjust the scope by removing certain elements from the constituent.

In step 1, the method parses the sentence and considers all the constituents that dominate the negation cue as scope candidates. In the sentence shown in Figure 1, scope candidates for the cue are RB, VP and S.

In step 2, the method selects one constituent from the candidates using the path pattern rules. In the example shown in Figure 1, the rule set shown in Figure 3 selects the constituent S. We add new path pattern rules to Yoshida et al. (2023)'s rules. The additional rules are based on the preliminary experiment using the training data, where we observed that Yoshida et al.'s rules do not cover verb and noun negation cues such as "lack". If none of the rules are activated, we use *default scope* proposed by Read et al. For the detail of *default scope*, see Read et al. (2012).

Step 3 is the scope adjustment, where heuristic rules remove certain elements that are not included in the scope. We observed that some rules of Yoshida et al. are specialized for the literary domain like Sherlock. For example, noun phrases

Method	Token-level (%)			Scope-level (%)		
	Pre.	Rec.	F_1	Pre.	Rec.	F_1
Baseline (BERT) *	94.44	89.23	91.76	99.11	71.77	83.25
Combination (BERT)	94.74	87.23	90.83	98.57	66.61	79.48
Conversion+Combination (BERT)	93.84	92.43	**93.13**	98.91	74.21	84.79
Baseline (RoBERTa) *	92.08	90.44	91.24	**99.45**	58.60	73.74
Combination (RoBERTa)	93.58	87.29	90.32	99.19	58.44	73.53
Conversion+Combination (RoBERTa)	91.47	92.10	91.76	99.08	60.53	75.14
Khandelwal and Sawant, 2020 (BERT) *	–	–	92.36	–	–	–
Truong et al., 2022 (RoBERTa) – Baseline *	–	–	91.51	–	–	–
Truong et al., 2022 (RoBERTa) – CueNB	–	–	91.24	–	–	–
Wu and Sun, 2023 (BERT)	**95.12**	90.57	92.77	–	–	85.35
Wu and Sun, 2023 (RoBERTa)	94.54	91.24	92.85	–	–	**87.10**
Yoshida et al., 2023 (heuristics) **	89.32	**94.30**	91.74	98.94	**74.70**	85.13

Table 2: Results for negation scope resolution. The scores in our experiments are an average of five runs with different random seeds. The previous studies used only Sherlock for the training. Values marked in bold and underlined represent the highest and the second-highest scores for each evaluation metric, respectively. * Baseline (BERT) and Baseline (RoBERTa) can be regarded as reproductions of Khandelwal and Sawant (2020) and Truong et al. (2022) – Baseline, respectively. The score difference between the reproduction and the original paper is due to the differences in random seeds. ** Yoshida et al. (2023) used PLMs only for syntactic parsing. Yoshida et al. (2023) reported only token-level F_1 in the exact evaluation metric as the others; hence we derived the other scores by reproduction.

delimited by punctuations, which are commonly found in dialogues in literary stories, are often excluded from the scope. Yoshida et al.'s rule removes punctuation-delimited noun phrases from the scope. However, in biological and review domains, noun phrases delimited by punctuations are frequently included in the scope. To address such a difference in domains, we use only a subset of Yoshida et al.'s rules that can apply to general text. To be specific, we conduct the following adjustments:

- Remove constituent-initial punctuations, RB, CC, UH, ADVP, INTJ or SBAR.

- Remove constituent-initial PP that is delimited by a punctuation.

- Remove punctuation-delimited ADVP or INTJ.

- Remove CC and previous conjuncts if the cue is in a conjoined phrase.

In the example shown in Figure 1, the rules remove "surprisingly" and the comma following it.

As seen in the example shown in Figure 1, we can accurately convert the scope of B&S to that of Sherlock.

4. Negation Scope Resolution with a Unified Dataset

We conducted experiments of negation scope resolution to evaluate the effectiveness of the proposed method in terms of scaling up the dataset

through scope conversion. We merged the training data of Sherlock and the converted version of B&S into a unified dataset. We used the unified dataset to fine-tune PLM-based models and performed negation scope resolution.

4.1. Experimental Settings

Since B&S does not have official data splits, we performed the data split for B&S in the same way Truong et al. (2022) did[8]. SFU and Sherlock are PTB-tokenized, whereas BioScope is not. As a pre-processing, we applied NLTK (Bird and Loper, 2004) tokenizer for BioScope.

We fine-tuned BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019)[9] for negation scope resolution using the code of Truong et al. (2022). Fine-tuning was conducted using three different configurations: (i) **Baseline**, using only the training data of Sherlock; (ii) **Combination**, using the training data of Sherlock and the unconverted (original) data of B&S; (iii) **Conversion+Combination (C+C)**, using the training data of Sherlock and the converted version of B&S. The parsing process was performed by Berkeley Neural Parser (Kitaev and Klein, 2018; Kitaev et al., 2019) with BERT (Devlin et al., 2019). For the validation and the evaluation of the models, we used the validation and test sets of Sherlock,

---

[8]We referred to https://github.com/joey234/negation-focused-pretraining.

[9]We used `bert-base-uncased` for BERT and `roberta-base` for RoBERTa, both of which are released in Hugging Face (https://huggingface.co/).

respectively. For all the settings, we adopted the same hyperparameters as those used by Truong et al. (2022).

4.2. Evaluation Metrics

We adopted the two evaluation metrics: token-level and scope-level measures, both of which compute precision, recall and F_1 measure. The token-level measure checks whether each token is correctly predicted, while the scope-level measure checks whether all tokens inside the scope are correctly predicted. A high scope-level score means a high capability for the model to predict scopes perfectly. The token-level measure is less strict, and a high token-level score means the overall little gap between the predicted scopes and the ground truth. Token-level and scope-level measures correspond to *Scope Tokens* and *Scopes CM* used in *SEM2012 shared task (Morante and Blanco, 2012). There are two differences from the metrics used in the shared task: punctuations are also included in the evaluation, and gold cues are given.

We adopted these two metrics because they are major ones that have been used in many previous studies since their proposal in *SEM2012 shared task. Using these metrics enables a fair comparison between the results of this study and those of the previous studies.

4.3. Experimental Results

Table 2 shows the experimental results. As shown in the previous studies, the simple combination of the three corpora (Combination) led to lower performance of the models. However, when using the unified dataset (C+C), the performance of the models improved. This supports the effectiveness of our conversion method in terms of scaling up the dataset through scope conversion. In particular, fine-tuning on the unified dataset mitigated the tendency for lower recall observed in the previous studies. For token-level F_1 measure, C+C with BERT outperformed state-of-the-art models: Wu and Sun (2023) and Truong et al. (2022).

5. Conclusion

This paper proposed a method for automatically converting the negation scopes of B&S to those of Sherlock and merged them into a unified dataset. To verify the effectiveness of the proposed method, we conducted experiments of negation scope resolution using the unified dataset for fine-tuning PLM-based models. The experimental results showed that a simple combination of the corpora resulted in lower performances of the models. However, when using our unified data for fine-tuning, the model performances improved, which supports

the effectiveness of the proposed method in terms of scaling up the dataset through scope conversion.

For future work, we plan to manually annotate B&S according to the guidelines of Sherlock to evaluate how accurate the proposed conversion method is. In addition, potential future work includes fine-tuning the state-of-the-art negation scope resolution model proposed by Wu and Sun (2023) on our unified dataset, in which we can expect to achieve better performance.

6. Limitations

The proposed method has two main limitations: (i) Parsing errors in step 1 lead to conversion errors. Although we adopted the high-accuracy parser (with 0.958 F_1 on Penn Treebank (Marcus et al., 1993)), parse errors still occur. (ii) The rules in the scope resolution steps 2 and 3 cannot cover all possible patterns.

7. Acknowledgements

8. Bibliographical References

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

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota. 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, pages 58–63, Valencia, Spain. Association for Computational Linguistics.

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.

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.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

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 8th International Conference on Language Resources and Evaluation*, pages 3190–3195, Istanbul, Turkey. European Language Resources Association.

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 pre-training approach.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Andrew Moore and Jeremy Barnes. 2021. Multi-task learning of negation and speculation for targeted sentiment classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2838–2869, Online. Association for Computational Linguistics.

Roser Morante and Eduardo Blanco. 2012. *SEM 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics*, pages 265–274, Montréal, Canada. Association for Computational Linguistics.

Roser Morante and Walter Daelemans. 2012. ConanDoyle-neg: Annotation of negation cues and their scope in conan doyle stories. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 1563–1568, Istanbul, Turkey. European Language Resources Association.

Roser Morante, Sara Schrauwen, and Walter Daelemans. 2011. Annotation of negation cues and their scope : Guidelines v1.0. Technical report, University of Antwerp.

Partha Mukherjee, Youakim Badr, Shreyesh Doppalapudi, Satish M. Srinivasan, Raghvinder S. Sangwan, and Rahul Sharma. 2021. Effect of negation in sentences on sentiment analysis and polarity detection. *Procedia Computer Science*, 185:370–379.

Jonathon Read, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2012. UiO1: Constituent-based discriminative ranking for negation resolution. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics*, pages 310–318, Montréal, Canada. Association for Computational Linguistics.

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 38–45, Columbus, Ohio. Association for Computational Linguistics.

Thinh Truong, Timothy Baldwin, Trevor Cohn, and Karin Verspoor. 2022. Improving negation detection with negation-focused pre-training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4188–4193, Seattle, United States. Association for Computational Linguistics.

Yin Wu and Aixin Sun. 2023. Negation scope refinement via boundary shift loss. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6090–6099, Toronto, Canada. Association for Computational Linguistics.

Asahi Yoshida, Yoshihide Kato, and Shigeki Matsubara. 2023. Revisiting syntax-based ap-

proach in negation scope resolution. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics*, pages 18–23, Toronto, Canada. Association for Computational Linguistics.