General Paper Peer-Reviewed

Negation Scope Conversion for Unifying Negation-Annotated Datasets

Asahi Yoshida[a], Yoshihide Kato[b] and Shigeki Matsubara[a,c]

Negation scope resolution is a technique that identifies the part of a sentence affected by the negation cue. The three major corpora used for it, the BioScope corpus, the SFU review corpus, and the Sherlock dataset, have different annotation schemes for negation scope. Due to the different annotations, it is difficult to use the three corpora together in the study of negation scope resolution. To address this issue by merging the corpora into a unified dataset based on a common annotation scheme, we propose a method for automatically converting the scopes of BioScope and SFU to those of Sherlock. We conducted an experiment to evaluate the accuracy of our method using a dataset obtained by manually annotating the negation scopes to a tiny portion of BioScope and SFU, verifying that our method can convert the scopes with high accuracy. In addition, we conducted another experiment to verify the effectiveness of our method from a pragmatic perspective, where we fine-tuned PLM-based negation scope resolution models using the unified dataset obtained by our method. The results demonstrated that the performances of the models increase when fine-tuned on the unified dataset, unlike the simply combined one, which supports the effectiveness of our method.

Key Words: *Negation, Negation Scope Conversion, Negation Scope Resolution*

1 Introduction

Negation is a frequently occurring linguistic phenomenon in natural languages. Accurate negation processing is crucial for natural language processing (NLP) systems because it reverses the meaning of sentences, phrases, words, etc. Over the past decade, many researchers in the NLP community have addressed the work on negation processing. Despite these extensive efforts,

[a] Graduate School of Informatics, Nagoya University
[b] Information & Communications, Nagoya University
[c] Information Technology Center, Nagoya University
This paper is an extended version of our preliminary paper (Yoshida et al. 2024), which was presented at the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).

recent studies have shown that pre-trained language models (PLMs) still struggle with negation (Hossain et al. 2022; Chen et al. 2023; Truong et al. 2023; Ye et al. 2023; García-Ferrero et al. 2023; Weller et al. 2024). The negation handling ability of language models has been evaluated from various perspectives, and some studies have pointed out that language models are not robust to variations in negation scope, which is the part of a sentence affected by the negation (Ravichander et al. 2022; She et al. 2023). Considering this point, this study addresses the fundamental technique of negation modeling, **negation scope resolution**: resolving the part(s) of a sentence affected by the negation cue (words that express the meaning of negation such as "not" and "no").

Many researchers have addressed negation scope resolution using various approaches; however, this remains a difficult problem due to its complexity. When using negation-annotated corpora for the study of negation scope resolution, one issue arises: The three primary negation-annotated corpora, the BioScope corpus (Szarvas et al. 2008), the SFU review corpus (Konstantinova et al. 2012), and the Sherlock dataset (Morante and Daelemans 2012), adopt different annotation schemes for the negation scope. Table 1 shows different annotations of negation cue and scope for the sentence "The park allows fishing in designated areas only and does not allow swimming.". Due to this difference in scope annotation, it is difficult to use the three corpora together in the study of negation scope resolution. For example, previous studies that performed negation scope resolution using PLMs trained a model on a single corpus instead of on a combined dataset consisting of the three corpora (e.g., Truong et al. (2022)), because a simple combination of the three corpora can decrease the model performance due to the different annotations of the negation scope. This issue can be resolved by merging the three corpora into a unified dataset based on a common annotation scheme. 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 annotations require expert linguists and considerable time and effort. Automated conversion of negation scopes can be an alternative solution. However, to the best of our knowledge, no previous study has developed such a method.

Corpus	Annotation of negation
BioScope	The park allows fishing in designated areas only and does **not** allow swimming.
SFU	The park allows fishing in designated areas only and does **not** allow swimming.
Sherlock	The park allows fishing in designated areas only and does **not** allow swimming.

Table 1 Negation cue and its scope for the example sentence according to the three annotation schemes. The bold and underlined parts represent the negation cue and scope, respectively.

Inspired by these issues, this study proposes a method for automatically converting the scopes of BioScope and SFU to those of Sherlock.[1] We select the Sherlock scope annotation as the target for conversion because it can represent more complex negation scopes than BioScope and SFU. The key point of our method is to effectively convert scopes by utilizing the existing scope annotations in BioScope and SFU.

An experiment was conducted to evaluate the accuracy of the proposed method. We obtained a dataset by manually annotating the negation scopes to a tiny portion of BioScope and SFU following the Sherlock standard. Using this dataset, we experimentally evaluated the accuracy of our method and performed a qualitative error analysis. The results demonstrated that our method can convert the scopes of BioScope and SFU to those of Sherlock with high accuracy.

Additionally, we conducted another experiment to verify the effectiveness of the proposed method from a pragmatic perspective. We converted BioScope and SFU using our method, and merged Sherlock and the converted versions of BioScope and SFU into a unified dataset. The unified dataset can be treated as a large training dataset for negation scope resolution models. We fine-tuned the PLMs on the unified dataset and evaluated the model performances. The experimental results demonstrated that the dataset created using our method improves the performance of the scope resolution model, unlike the simply combined one.

The remainder of this paper is organized as follows: Section 2 describes the related work of our study. Section 3 proposes a method for automatically converting the negation scopes of BioScope and SFU to those of Sherlock. Section 4 reports an experiment demonstrating that our method can convert the scopes of BioScope and SFU to those of Sherlock with high accuracy. Section 5 reports an experiment of negation scope resolution using the unified dataset obtained by the conversion method, demonstrating the better performance of the model trained on the unified dataset than the one on the simply combined dataset. Finally, Section 6 concludes the paper.

2 Related Work

2.1 Negation-Annotated Corpora

This section describes the three negation-annotated corpora used in this study.

[1] Our code is available at `https://github.com/asahi-y/negation-scope-conversion`. The unified dataset can be obtained by applying our code to BioScope and SFU, and merging them with Sherlock.

BioScope The BioScope corpus (Szarvas et al. 2008)[2] contains biological texts in which negation cues are annotated with their scopes. Sentences (1) and (2) show examples annotated with negation cues (marked in bold) and their scopes (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 side of the cue. Exceptional cases exist in which the scope annotation also covers the left side 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 that belong to 8 different domains such as books, cars, and hotels. 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), (4), and (5) present annotation examples.

(3) We did **not** drive up to the door.
(4) You'll see how **im**possible for me to go there.
(5) An investigator needs facts and **not** legends or rumours.

The annotation guidelines of Sherlock (Morante et al. 2011) are based on those of BioScope; however, several improvements have been made. 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 (e.g., an elliptical construction is such a case, as shown in Sentence (5)).

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

[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

2.2 Negation Scope Resolution

This section explains the negation scope resolution methods used as the basis for our conversion method. We selected these methods due to their high performance.

2.2.1 Syntax-Based Negation Scope Resolution

One method that we adopt is the syntax-based negation scope resolution method proposed by Yoshida et al. (2023). They modified the syntax-based method proposed by Read et al. (2012). Yoshida et al.'s method follows these steps:

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.

As an example, let us consider Sentence (5) with a negation cue.

(5) He did **not** go to school and stayed home.

In Step 1, the method obtains the parse tree, as shown in Figure 1, and selects the RB, two VPs, and S as the negation scope candidates. In Step 2, the constituent S is selected using the path pattern rules. Step 3 removes the coordinating conjunction "and," its conjunct "stayed home," and the punctuation "." according to the adjustment rules (we explain the details of the new version of the path pattern rules and the adjustment rules in the next section). Using a series of processes, the scope "He did, go to school" is correctly resolved.

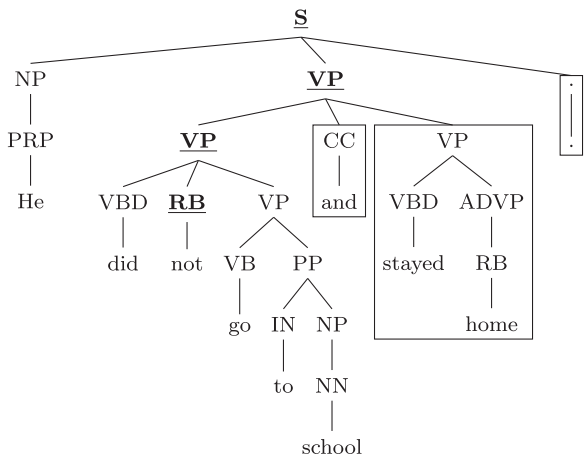

Figure 1 Parse tree of Sentence (5). Underlined constituents show scope candidates. Enclosed parts are removed in the adjustment step.

Yoshida et al. (2023) demonstrated that by using a modern parser, the scope resolution method based on heuristics for syntax can obtain the same level of performance as the state-of-the-art methods based on end-to-end neural networks.

2.2.2 NegBERT

The other method that we adopt is called NegBERT, which was proposed by Khandelwal and Sawant (2020). This method applies the transfer learning approach of PLMs to a negation scope resolution task. As a base model, this method uses BERT (Devlin et al. 2019) with a classification layer on top of it. NegBERT can be obtained by fine-tuning this model on a negation-annotated corpus. This method models the negation scope resolution task as an end-to-end binary sequence labeling task. It encodes the sentence along with the position and the type of a negation cue and then, the model outputs whether each word in the sentence is within the negation scope.

3 Negation Scope Conversion

This section proposes a method for automatically converting the negation scopes of BioScope[5] and SFU to those of Sherlock. As explained in Section 2.1, most annotation schemes of the negation scope in SFU follow those of BioScope. Therefore, we regard the scope annotations of BioScope and SFU as almost identical and treat them in the same way.[6] We refer to the combination of BioScope and SFU as B&S. Below, we represent a negation scope as a set of indices, each indicating the position of a word in a sentence.

The key idea of our method is to convert scopes effectively by utilizing the existing scope annotations in B&S. This idea is based on the following hypotheses, which were developed by observing the training data and the annotation guidelines of the three corpora:

1. The right parts of the scope to the cue can be regarded as almost identical for B&S and Sherlock.
2. The middle parts of the scope in the cue can be regarded as almost identical for B&S and Sherlock.
3. The left parts of the scope to the cue can be regarded as almost identical for B&S and Sherlock if the scope of B&S covers the left of the cue (e.g., Sentence (2) is the case).

[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) are 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.

Based on these hypotheses, we established the following strategy for conversion:

- We divide the negation scope into four parts: the left part (denoted as S_{left}), the middle part (S_{mid}), the right part (S_{right}), and the internal part of the cue (S_{cue}).
- For S_{mid} and S_{right}, we utilize the existing scope annotations in B&S, following Hypotheses 1 and 2.
- For S_{left}, we utilize the existing scope annotations in B&S if the scope annotation in B&S covered the left part of the cue, following Hypothesis 3. If the scope annotation in B&S did not cover the left part of the cue, we obtain the left part of the scope using the negation scope resolution method.

Below is the formal definition of our conversion method that obtains the converted scope $S_{\text{converted}}$:

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

where

$$S_{\text{left}} = \begin{cases} \mathrm{L}_c(S_{\text{B\&S}}) & (\mathrm{L}_c(S_{\text{B\&S}}) \neq \emptyset) \\ \mathrm{L}_c(S_{\text{res}}) & (\mathrm{L}_c(S_{\text{B\&S}}) = \emptyset), \end{cases}$$

$$S_{\text{mid}} = \mathrm{M}_c(S_{\text{B\&S}}),$$

$$S_{\text{right}} = \mathrm{R}_c(S_{\text{B\&S}}).$$

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

$$\mathrm{L}_c(S) = \{i \in S \mid i < c_l\},$$
$$\mathrm{M}_c(S) = \{i \in S \mid c_l < i < c_r\},$$
$$\mathrm{R}_c(S) = \{i \in S \mid c_r < i\}.$$

Here, c_l and c_r denote indices for the leftmost and rightmost words of the negation cues, respectively.

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

3.1 Affixal and Contracted Cues

Sherlock distinguishes the cue and its scope in the word for affixal cues (e.g., **un**usual) and contracted cues (e.g., do**n't**), whereas B&S treats the entire word as the cue. We distinguish affixal and contracted cues in B&S based on simple pattern matching. If B&S's cue c has an affix in $\mathrm{V}_{\mathrm{aff}}$ or a suffix in $\mathrm{V}_{\mathrm{cont}}$, S_{cue} is a singleton set consisting of the result by removing the affix from the cue. Based on the annotation guidelines of Sherlock (Morante et al. 2011) and the training data, we define $\mathrm{V}_{\mathrm{aff}}$ and $\mathrm{V}_{\mathrm{cont}}$ as follows:

$$\mathrm{V}_{\mathrm{aff}} = \{\mathrm{dis}, \mathrm{im}, \mathrm{in}, \mathrm{ir}, \mathrm{un}, \mathrm{less}\} \ , \ \mathrm{V}_{\mathrm{cont}} = \{\mathrm{n't}, \mathrm{not}\} \ .$$

Figure 2 shows an example of a scope conversion that includes a contracted cue.

3.2 Negation Scope Resolution

To obtain S_{res}, we can adopt any scope resolution method. In the experiments reported later, we chose to use the methods described in Section 2.2.

For the method proposed by Yoshida et al. (2023), we modify Steps 2 and 3 explained in Section 2.2.1. The modifications adapt to the domains of BioScope and SFU. Below, we explain our modifications using the example shown in Figure 3 and the parse tree shown in Figure 4. In Step 2, we add new path pattern rules to Yoshida et al.'s rules. The additional rules are based on a 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." Figure 5 shows the new rule

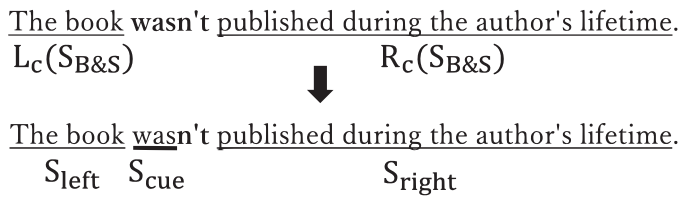

Figure 2 Example of scope conversion where $\mathrm{L}_c(S_{\mathrm{B\&S}}) \neq \emptyset$ and a contracted cue is included.

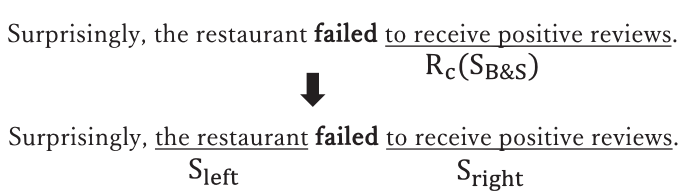

Figure 3 Example of scope conversion where $\mathrm{L}_c(S_{\mathrm{B\&S}}) = \emptyset$.

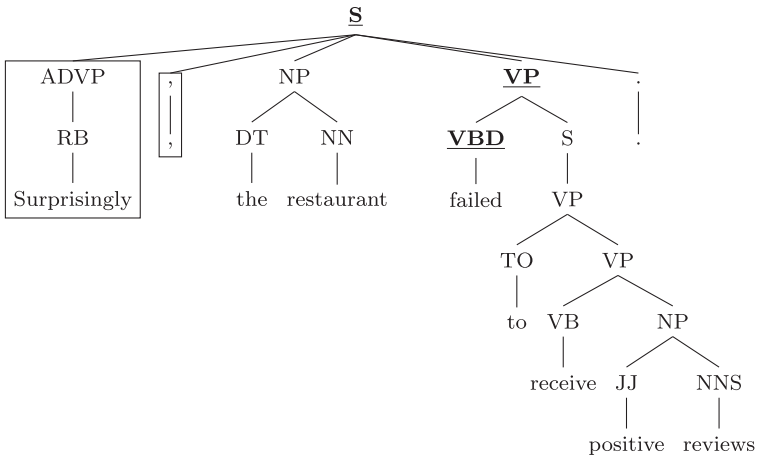

Figure 4 Parse tree of the sentence "Surprisingly, the restaurant failed to receive positive reviews.". Underlined constituents denote scope candidates. Enclosed parts are removed in the adjustment step.

```
RB//VP/S/SBAR if SBAR\WHNP
RB//VP/S
RB//S
DT/NP if NP/PP
DT//SBAR if SBAR\WHADVP
DT//S
JJ//ADJP/VP/S if S\VP\VB* [@lemma="be"]
JJ/NP/NP if NP\PP
JJ//NP
```

```
UH
IN/PP
NN//NP/NP if lemma of NN in ["lack", "absence"]
NN/NP//S/SBAR if SBAR\WHNP
NN/NP//S
CC/SINV
VBG//NP
VBN//NP
VB*//S
```

Figure 5 Path pattern rules. Each row represents one rule, shown in the order in which they are applied. X/Y means Y is the parent of X, X//Y denotes Y is an ancestor of X, and X\Y implies Y is a child of X. The symbol * matches any sequence of zero or more characters, namely, VB* denotes any string starting with VB. The underlined rules are those that we add in this study.

sets.[8] In the example shown in Figure 3, the rule set selects the constituent S that dominates the entire sentence, with the additional rule `VB*//S` activated. If none of the rules are activated, we use *default scope* proposed by Read et al. For the details on *default scope*, see Read et al. (2012). For Step 3, we modify Yoshida et al.'s scope adjustment rules. We observed that some of Yoshida et al.'s rules specialize in the literary domain, such as Sherlock. For example, noun

[8] As supplementary information, we show the frequency of each rule's application on the training sets in Appendix A.

phrases delimited by punctuations, which are commonly found in dialogues in literary stories, are often excluded from the scope. Based on this, Yoshida et al.'s rule removes punctuation-delimited noun phrases from the scope. However, in the biological and review domains, noun phrases delimited by punctuations are frequently included within 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, our method performs 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 3, the rules remove "surprisingly" and the comma following it.

As seen in the examples shown in Figures 2 and 3, we can accurately convert the scope of B&S to that of Sherlock.

4 Evaluation of the Accuracy of the Conversion Method

We evaluated how accurately the proposed method converts the scopes of B&S to those of Sherlock. Section 4.1 describes the evaluation metrics used to evaluate the accuracy of the conversion method. For the evaluation, we conducted an experiment as follows: First, we obtained a dataset by manually annotating the negation scopes to a tiny portion of B&S, namely, the test and validation sets of B&S, following the annotation guidelines of the Sherlock dataset (Section 4.2). Using this dataset, we evaluated our method based on the metrics (Section 4.3). We also qualitatively analyzed the results by performing a manual error analysis on the validation set (Section 4.4).

4.1 Evaluation Metrics

This section describes the evaluation metrics used to evaluate the accuracy of the conversion method. In the evaluation, we focus specifically on the important hypotheses, which the method adopts: (i) $\mathrm{R}_c(S_{\mathrm{B\&S}})$ and $\mathrm{R}_c(S_{\mathrm{SH}})$ can be regarded as almost identical, (ii) $\mathrm{M}_c(S_{\mathrm{B\&S}})$ and $\mathrm{M}_c(S_{\mathrm{SH}})$ can be regarded as almost identical, and (iii) $\mathrm{L}_c(S_{\mathrm{B\&S}})$ and $\mathrm{L}_c(S_{\mathrm{SH}})$ can be regarded as almost identical if $\mathrm{L}_c(S_{\mathrm{B\&S}}) \neq \emptyset$. The evaluation metric ***hypo-token*** considering this point is summarized as follows:

- *Hypo-token* checks whether the converted scopes and the ground-truth match at the token level.

- *Hypo-token* evaluates only the tokens that belong to the scopes that were assumed to be almost identical between B&S and Sherlock. Specifically, the tokens that belong to the following parts are evaluated:
  - $\mathrm{R}_c(S_{\text{converted}})$ and $\mathrm{M}_c(S_{\text{converted}})$.
  - $\mathrm{L}_c(S_{\text{converted}})$ if $\mathrm{L}_c(S_{\text{B\&S}}) \neq \emptyset$.

We adopted *hypo-token* as the primary evaluation metric.

In addition to *hypo-token*, we used the following two metrics that are commonly used for the evaluation of negation scope resolution:

- ***Scope-strict-match***[9]: This metric checks the scope labels for all the tokens inside the scope for each negation cue; TP, TN, FP, or FN is assigned to each scope. To consider the scope as TP, all tokens inside the scope of a negation cue must match between the target and the ground-truth. Partial matches are considered as FN. A high score in this metric indicates the high capability of the method (or model) to convert (or predict) scopes perfectly.
- ***Scope-token***: This metric checks the scope label of each token; TP, TN, FP, or FN is assigned to each token. A high score on this metric indicates an overall small gap between the two scopes.

For all the metrics, the precision, recall, and F_1 measures are calculated. If a sentence contains more than one negation cue, evaluation is conducted for each cue separately. Note that punctuation tokens are also included in the evaluation.

4.2 Manual Scope Annotation

Since the proposed method converts the scopes of B&S to those of Sherlock, the evaluation requires B&S data annotated with negation scopes following the Sherlock standard. We constructed a dataset by manually annotating a tiny portion of B&S, namely the test and validation sets[10] with negation scopes following the annotation guidelines of Sherlock (Morante et al. 2011). The test set was used to measure the accuracy of the conversion method and the validation set was used to measure the inter-annotator agreement (IAA) and manual error analysis. We applied the NLTK (Bird and Loper 2004) tokenizer to BioScope as a pre-processing step because SFU and Sherlock are tokenized based on the Penn Treebank (Marcus et al. 1993) standard, whereas BioScope is not.

[9] This metric is commonly referred to as *scope(s)* or *scope-level*; however, in this paper, we refer to it as *scope-strict-match* for clarity.

[10] The train-val-test split is identical to that of Truong et al. (2022). We referred to `https://github.com/joey234/negation-focused-pretraining`.

4.2.1 Annotation Details

The annotation was performed by a worker (not one of the authors) who was proficient in the NLP annotation tasks. We asked the worker to annotate the negation scopes following the Sherlock standard, given the sentences and negation cues of B&S. With regard to the negation cues, we instructed the worker to use cues as they are; however, one exception exists: The worker extracted substrings from the negation cues of B&S, if they are affixal or contracted cues in the Sherlock standard. This is intended to handle the difference in the annotations for affixal and contracted cues between B&S and Sherlock as mentioned in Section 3.1.

The annotation process is as follows:

1. The authors provided the worker with instructions for the annotation task and the annotation guidelines of Sherlock (Morante et al. 2011). A small set of instances[11] and their correct annotations were also provided to facilitate the worker's understanding of the task. Using these materials, the worker understood the details of the annotation process.
2. The authors provided the worker with the test and validation sets of B&S and the worker performed the annotation task. Each instance contained a sentence and a negation cue. Note that the worker was not provided with the scope annotation of B&S. It should also be noted that the worker and researchers did not discuss the content of scope annotations.

4.2.2 Annotation Quality Assessment

To assess the quality of annotations by the worker, one of the authors annotated a subset of the validation sets: a total of 100 sentences, consisting of 50 randomly extracted sentences from each corpus (out of 291 sentences in the BioScope validation set and 468 of SFU). For the 100 sentences, we calculated the IAA between the worker and the author. Below, the scope annotations by the worker and the author are called $\mathrm{SA}_{\mathrm{worker}}$ and $\mathrm{SA}_{\mathrm{author}}$, respectively.

We calculated the IAA in terms of precision, recall, and F_1 measure following Morante and Daelemans (2012). For the calculation, $\mathrm{SA}_{\mathrm{author}}$ was assessed against $\mathrm{SA}_{\mathrm{worker}}$. IAA was calculated by *scope-strict-match* and *scope-token*, both of which are explained in Section 4.1.

Tables 2 and 3 show the IAA scores between $\mathrm{SA}_{\mathrm{worker}}$ and $\mathrm{SA}_{\mathrm{author}}$. For comparison, we present the IAA scores in the scope annotations of Sherlock in Table 4, obtained by Morante and Daelemans (2012). At the metric of *scope-strict-match*, the IAA score between $\mathrm{SA}_{\mathrm{worker}}$ and $\mathrm{SA}_{\mathrm{author}}$ had a precision of 100% and a recall of approximately 75% (Table 2). This implies approximately 75% of the scope annotations between the worker and the author match perfectly,

[11] The instances were extracted from the training set of B&S.

Corpus	Pre.	Rec.	F_1	#Scopes	TP	FP	FN
BioScope (50 sentences in the validation set)	100	74.1	85.1	58	43	0	15
SFU (50 sentences in the validation set)	100	75.5	86.0	53	40	0	13
All (100 sentences in the validation sets)	100	74.8	85.6	111	83	0	28

Table 2 IAA in *scope-strict-match*: scope annotations by the author was assessed against those by the worker.

Corpus	Pre.	Rec.	F_1	#Tokens	TP	FP	FN
BioScope (50 sentences in the validation set)	96.7	83.7	89.7	1790	701	24	137
SFU (50 sentences in the validation set)	97.8	84.9	90.9	1256	395	9	70
All (100 sentences in the validation sets)	97.1	84.1	90.1	3046	1096	33	207

Table 3 IAA in *scope-token*: scope annotations by the author was assessed against those by the worker.

whereas approximately 25% differ partially, considering that TP represents perfect matches and FN represents partial matches in the metric of *scope-strict-match*. Considering the recall of 79.08% and 69.43% in the Sherlock for the *cue-scope-strict* metric, the scores between SA_{worker} and SA_{author} can be comparable, whereas we cannot directly compare the *cue-scope-strict* and *scope-strict* metrics. For the *scope-token* metric, the IAA score between SA_{worker} and SA_{author} has an F_1 measure of approximately 90% (Table 3), which is comparable to the scores for the Sherlock: 91.53% and 88.04% (Table 4).[12] Considering these points, we concluded that the quality of SA_{worker} was acceptable and used it to evaluate the conversion method.

4.3 Experiment

We experimentally evaluated the accuracy of the proposed conversion method using the dataset described in Section 4.2. We applied the conversion method to the test sets of B&S to obtain its converted version, which we refer to as $SA_{converted}$. $SA_{converted}$ was evaluated against SA_{worker}. For the scope resolution process in the conversion method, we used two methods: Yoshida et al. (2023)'s one with modifications and Khandelwal and Sawant (2020)'s one (NegBERT), both of which are explained in Section 2.2. The parsing process in the method proposed by Yoshida et al.'s was performed using the Berkeley Neural Parser (Kitaev and Klein 2018; Kitaev et al. 2019) with BERT (Devlin et al. 2019). For obtaining NegBERT, we fine-tuned

[12] From Tables 3 and 4, we can observe that for the metric of *scope-token*, the IAA scores in this study have higher precision than recall, whereas those in the Sherlock dataset have higher recall than precision. This difference is due to the selection of the ground-truth in the IAA calculation. The precision and recall values vary depending on which annotator's annotation is used as the ground-truth for the agreement calculation.

Dataset	Metric	Pre.	Rec.	F_1
Hound (training set)	*Cue*	94.02	95.76	94.88
	Cue-scope-strict	91.98	79.08	85.04
	Scope-token	89.95	93.17	91.53
Wisteria (validation set)	*Cue*	90.59	95.06	92.77
	Cue-scope-strict	87.20	69.43	77.31
	Scope-token	85.88	90.31	88.04

Table 4 IAA in the Sherlock dataset. The scores show the IAA between two annotators, quoted from Morante and Daelemans (2012). The annotations of one annotator were assessed against those of the other annotator (official annotation of Sherlock). Since Sherlock annotates both negation cues and their scopes, the IAA metrics are partially different between Morante and Daelemans (2012), and those of this paper. *Cue* evaluates whether tokens of negation cues match; *cue-scope-strict* (referred to as *scopes* in the original paper) evaluates whether negation cues and their scopes match strictly; *scope-token* is the same metric as this paper. Note that *cue-scope-strict* is a stricter metric than *scope-strict* in this paper.

BERT on the Sherlock dataset for negation scope resolution using the code and the hyperparameters provided by Truong et al. (2022).[13]

The accuracy of the conversion method was evaluated based on the three metrics described in Section 4.1. As a comparison method, we directly used the scope resolution methods, which simply predict the scope labels for all tokens without using any scope labels of B&S. This can be considered as an ablation of our conversion.

Table 5 shows the evaluation results of the *hypo-token* metric. For this metric, the proposed conversion method achieved an F_1 measures of 89.95% and 82.08% for BioScope and SFU, respectively. The method achieved relatively high precision scores for both BioScope and SFU. Particularly, the score in BioScope exceeded 98%. This indicates that almost all tokens within the scopes of BioScope are also included in Sherlock scopes for the parts where we assumed that the scopes are almost identical between B&S and Sherlock. The precision for SFU also exceeds 90%, confirming that most of the tokens within the scopes of SFU are also included in those of Sherlock for the mentioned parts.

For recall, in contrast, the method shows relatively lower scores: 82.59% for BioScope and 75.00% for SFU. This shows that approximately 17.41% of the tokens in BioScope and 25.00% in SFU are not included in the scopes of B&S but are included in those of Sherlock for the parts where we assumed that scopes are almost identical between B&S and Sherlock. To identify the

[13] We referred to `https://github.com/joey234/negation-focused-pretraining`. This code includes the re-implementation of Khandelwal and Sawant (2020), which we used for obtaining NegBERT.

Corpus	Part	Pre.	Rec.	F_1	#Tokens	TP	FP	FN
BioScope	$L_c(S_{\text{converted}})$	98.72	83.05	90.21	724	387	5	79
	$M_c(S_{\text{converted}})$	100.00	100.00	100.00	7	7	0	0
	$R_c(S_{\text{converted}})$	98.76	82.44	89.87	4191	1916	24	408
	All	98.76	82.59	89.95	4922	2310	29	487
SFU	$L_c(S_{\text{converted}})$	0.00	N/A	N/A	11	0	1	0
	$M_c(S_{\text{converted}})$	N/A	N/A	N/A	0	0	0	0
	$R_c(S_{\text{converted}})$	90.68	75.00	82.10	6642	2568	264	856
	All	90.65	75.00	82.08	6653	2568	265	856
All	$L_c(S_{\text{converted}})$	98.47	83.05	90.10	735	387	6	79
	$M_c(S_{\text{converted}})$	100.00	100.00	100.00	7	7	0	0
	$R_c(S_{\text{converted}})$	93.96	78.01	85.25	10 833	4484	288	1264
	All	94.32	78.41	85.63	11 575	4878	294	1343

Table 5 Evaluation results for *hypo-token*: $SA_{\text{converted}}$ was evaluated against SA_{worker}. Along with the overall scores (All), scores for the left, middle, and right parts of the scopes ($L_c(S_{\text{converted}})$, $M_c(S_{\text{converted}})$, $R_c(S_{\text{converted}})$) are also provided. Note that for *hypo-token*, the score for $L_c(S_{\text{converted}})$ is counted only if $L_c(S_{\text{B\&S}}) \neq \emptyset$.

pattern of these cases, which our hypotheses do not cover, we perform a manual error analysis and provide a detailed discussion in Section 4.4.

Examining each part, the proportion of target tokens in $L_c(S_{\text{converted}})$ is relatively low (735 out of 11575). This is because, as mentioned in Section 2.1, scope annotations in B&S cover the left part of the cue only in exceptional cases, such as passive sentences. In these cases, the proposed method achieved an F_1 measure of 90.21% in BioScope. With regard to SFU, target tokens of $L_c(S_{\text{converted}})$ are only 11, which implies that exceptional cases rarely appear in the test set of SFU. The target tokens of $M_c(S_{\text{converted}})$ are also low because the middle part of the scope exists only in limited negation cues such as *neither, nor*. In this experiment, seven target tokens were present, all of which were accurately converted. We observe that the majority of the targets for *hypo-token* are in $R_c(S_{\text{converted}})$, that is, $R_c(S_{\text{converted}})$ is the most important in *hypo-token* to achieve a high accuracy of conversion. For this part, the performance of the conversion method was 89.87% for BioScope and 82.10% for SFU in terms of the F_1 measure. This relatively low score for SFU affects the overall score of SFU for *hypo-token*, making it relatively lower than that for BioScope. The reasons for this are discussed in Section 4.4.

Tables 6 and 7 show the evaluation results for the metrics of *scope-strict-match* and *scope-token*, respectively. For *scope-strict-match*, the scope conversion methods, both with Yoshida et al. (2023) and NegBERT, outperformed the scope resolution methods. In particular, scope

Method	Corpus	Pre.	Rec.	F_1	#Scopes	TP	FP	FN
Scope Conversion with Yoshida et al. (2023)	BioScope	99.51	66.01	79.37	307	202	1	104
	SFU	94.39	51.83	66.92	536	269	16	250
	All	96.52	57.09	71.74	843	471	17	354
Scope Conversion with NegBERT	BioScope	99.36	50.65	67.10	307	155	1	151
	SFU	100.00	48.21	65.06	536	272	13	248
	All	96.83	51.69	67.40	843	427	14	399
Scope Resolution by Yoshida et al. (2023)	BioScope	99.09	35.62	52.40	307	109	1	197
	SFU	91.71	38.42	54.15	536	199	18	319
	All	94.19	37.38	53.52	843	308	19	516
Scope Resolution by NegBERT	BioScope	99.19	39.87	56.88	307	122	1	184
	SFU	95.47	56.73	71.17	536	295	14	225
	All	96.53	50.48	66.30	843	417	15	409

Table 6 Evaluation results for *scope-strict-match*: $SA_{converted}$ was evaluated against SA_{worker}. Note that we applied the modifications explained in Section 3.2 to the method proposed by Yoshida et al. (2023).

Method	Corpus	Pre.	Rec.	F_1	#Tokens	TP	FP	FN
Scope Conversion with Yoshida et al. (2023)	BioScope	92.69	82.97	87.56	9463	3499	276	718
	SFU	80.91	74.64	77.65	12 619	3708	875	1260
	All	86.23	78.46	82.16	22 082	7207	1151	1978
Scope Conversion with NegBERT	BioScope	96.11	73.18	83.09	9463	3086	125	1131
	SFU	89.79	72.08	79.97	12 619	3581	407	1387
	All	92.61	72.59	81.38	22 082	6667	532	2518
Scope Resolution by Yoshida et al. (2023)	BioScope	81.99	92.72	87.02	9463	3910	859	307
	SFU	73.10	90.18	80.74	12 619	4480	1649	488
	All	76.99	91.34	83.55	22 082	8390	2508	795
Scope Resolution by NegBERT	BioScope	94.75	71.88	81.74	9463	3031	168	1186
	SFU	93.18	79.49	85.79	12 619	3949	289	1019
	All	93.86	75.99	83.99	22 082	6980	457	2205

Table 7 Evaluation results for *scope-token*: $SA_{converted}$ was evaluated against SA_{worker}. Note that we applied the modifications explained in Section 3.2 to the method proposed by Yoshida et al. (2023).

conversion with Yoshida et al. (2023) achieved significantly higher scores than scope resolution by Yoshida et al. (2023). This performance difference was mainly due to the gap in recall scores: 57.09% for conversion and 37.38% for resolution. A low recall score for *scope-strict-match* implies the number of partial matches is high compared with that of perfect matches. Therefore, the results suggest that our utilization of the B&S scopes is effective for scope conversion in increasing

the number of perfect scope matches. In terms of *scope-token*, in contrast, the conversion methods showed slightly lower scores than the resolution methods for the F_1 measure for both Yoshida et al. (2023) and NegBERT. The scores of the conversion methods show relatively higher precision and lower recall, which is consistent with the results for *hypo-token*, shown in Table 5. To identify the pattern of error cases, we report manual error analysis in the next section. As mentioned in Section 4.1, a *hypo-token* is a more appropriate metric than *scope-token* for evaluating the proposed conversion method.

Table 8 presents the evaluation results for the tokens predicted by the scope resolution method, that is, token-level evaluation for the scopes $L_c(S_{\text{converted}})$ where $L_c(S_{\text{B\&S}}) = \emptyset$, which is the subset of *scope-token*. Comparing the scores of Yoshida et al. (2023)'s method and NegBERT, a difference was observed between BioScope and SFU: The method proposed by Yoshida et al. achieved higher F_1 measures on BioScope than on SFU, whereas NegBERT achieved higher F_1 measures on SFU than on BioScope. This difference can be explained by the influence of both the domain of the corpus and the characteristics of the scope resolution methods. BioScope consists of formal, grammatically structured biomedical texts that allow the syntax-based method proposed by Yoshida et al. to resolve scopes more accurately than NegBERT, even with complex syntactic structures. In contrast, SFU consists of review texts that are often informal and non-grammatical, which makes it difficult for the syntax-based method to resolve scopes accurately.

Referring to Tables 5 and 8, we discuss the degree of the importance of the two factors in our method: the correctness of the hypotheses and the accuracy of the scope resolution method. Comparing the number of target tokens, the number of *hypo-token* is greater than that of tokens predicted by the scope resolution methods (11575 vs. 9509). This indicates that the hypotheses play a more important role than the scope resolution method in enhancing the accuracy of the

Resolution Method	Corpus	Pre.	Rec.	F_1	#Tokens	TP	FP	FN
Yoshida et al. (2023)	BioScope	82.63	83.57	83.10	4211	1175	246	231
	SFU	62.30	71.39	66.53	5298	1008	610	404
	All	71.81	77.47	74.53	9509	2183	857	635
NegBERT	BioScope	88.81	54.20	67.31	4211	762	96	644
	SFU	86.12	62.39	72.36	5298	881	142	531
	All	87.35	58.30	69.93	9509	1643	238	1175

Table 8 Evaluation results for the tokens predicted by the scope resolution methods. This is the subset of *scope-token* where $L_c(S_{\text{converted}})$ if $L_c(S_{\text{B\&S}}) = \emptyset$. Note that we applied the modifications explained in Section 3.2 to the method proposed by Yoshida et al. (2023).

proposed conversion method. However, since no significant difference exists in the number of target tokens, the performance of a negation scope resolution method is also important for the conversion method. Given that the performances of the scope resolution methods is relatively low (74.53%/69.93% for F_1 measure) compared with the score in *hypo-token* (85.63% for F_1 measure), we need a negation scope resolution method that can achieve higher accuracy across different domains in order to improve the accuracy of the conversion method.

4.4 Manual Error Analysis

We performed manual error analysis of the conversion method on the validation set of B&S, using the output of Scope Conversion with Yoshida et al. (2023), which outperformed Scope Conversion with NegBERT. We collected the instances in which the conversion output $S_{\text{converted}}$ did not match the scope annotation S_{worker}, namely, the instances where scope labels of $S_{\text{converted}}$ and S_{worker} differed in at least one token for the negation cue. Here, S_{worker} represents the scope annotation of the worker in each instance. The collected instances on the validation sets were 349 in total (98 in the BioScope Abstract sub-corpus, 32 in the BioScope FullPaper sub-corpus and 219 in SFU). Out of these, we extracted 230 instances (98 in BioScope Abstract, 32 in BioScope FullPaper and the initial 100 in SFU) as candidates for analysis and re-checked the scope annotation S_{worker}. We observed obvious mistakes in S_{worker} in 15 instances (8 in BioScope Abstract and 7 in SFU). We corrected these mistakes to remove their effects. If the correction led to $S_{\text{converted}}$ and S_{worker} matching, such instances were excluded. Finally, we manually analyzed 215 instances (90 in BioScope Abstract, 32 in BioScope FullPaper, and 93 in SFU). Conversion errors were observed in the left parts $\text{L}(S_{\text{converted}})$, or right parts $\text{R}(S_{\text{converted}})$ of the scopes; we counted the error patterns for $\text{L}(S_{\text{converted}})$ and $\text{R}(S_{\text{converted}})$ separately.

We classified the causes of the errors into the following categories:

- $\text{E}_{\text{hypotheses}}$: The hypotheses of the conversion method are invalid. In other words, the scope that was assumed to be almost identical between B&S and Sherlock is actually different.
- $\text{E}_{\text{resolution}}$: The prediction of the scope resolution method, used for the parts where the scopes of B&S and Sherlock were not assumed to be almost identical, is incorrect.
- $\text{E}_{\text{B\&S}}$: The scope annotation assigned to B&S is incorrect in the B&S standard.

Table 9 shows the classification statistics. The instances classified as $\text{E}_{\text{hypotheses}}$ are the most important ones for investigating the proposed conversion method. Some instances of errors in SFU classified as $\text{E}_{\text{hypotheses}}$ contained informal or ungrammatical expressions. The following instance is such an example:

(a) L($S_{\text{converted}}$)

Causes	BioScope Abstract	BioScope FullPaper	SFU
$E_{\text{hypotheses}}$	5	2	0
$E_{\text{resolution}}$	57	15	65
$E_{\text{B\&S}}$	0	4	0

(b) R($S_{\text{converted}}$)

Causes	BioScope Abstract	BioScope FullPaper	SFU
$E_{\text{hypotheses}}$	20 (6)	8 (4)	19 (5)
$E_{\text{B\&S}}$	19	8	20

Table 9 Number of error instances by causes for L($S_{\text{converted}}$) and R($S_{\text{converted}}$). The bracketed numbers in (b) represent the number of instances classified as $E_{\text{hypotheses}}$ that contain clauses or phrases expressing reason, which are related to the variations in the scope annotation following the guidelines of Sherlock. Note that the unbracketed value in each column includes the bracketed value.

- $S_{\text{converted}}$
  No *doubt, beautiful.*
- S_{worker}
  No *doubt, beautiful.*
- $S_{\text{B\&S}}$
  No *doubt, beautiful.*

In this example, $S_{\text{converted}}$ and S_{worker} differ in the right part of the cue: $S_{\text{converted}}$ included the adjective *beautiful* in the scope but S_{worker} did not. This instance includes informal expressions and is not a typical complete sentence with a subject and a verb. Although Sherlock provides detailed annotation guidelines to handle such incomplete sentences, the guidelines of BioScope and SFU do not mention such cases and likely rely on annotator judgment. This difference in the guidelines can cause different annotations between B&S and Sherlock for informal or incomplete sentences, which implies that our hypotheses do not hold in such cases. Since SFU consists of review-domain text, informal or ungrammatical expressions frequently appear as seen in the example above. In contrast, BioScope consists of formal biomedical texts, where such expressions are less frequent. Consequently, the lower *hypo-token* score in SFU compared with BioScope seen in Table 5 is partly due to the cases in which the hypotheses do not hold in informal or ungrammatical sentences. This is a limitation of the proposed method, which should be addressed in future work to enhance the conversion accuracy.

Moreover, $E_{\text{hypotheses}}$ has a frequent pattern in which a sentence contains adverbial clauses or phrases that express reason (such as a *because* clause). This pattern is explained using the following example:

- $S_{\text{converted}}$
  This shift is apparently ***not*** *caused by a recruitment phenomenon, because in FCS+ culture, the total number of colonies is not significantly modified by RA addition.*
- S_{worker}
  This shift is apparently ***not*** *caused by a recruitment phenomenon, because in FCS+ culture, the total number of colonies is not significantly modified by RA addition.*
- $S_{\text{B\&S}}$
  This shift is apparently ***not*** *caused by a recruitment phenomenon, because in FCS+ culture, the total number of colonies is not significantly modified by RA addition.*

In this instance, $S_{\text{converted}}$ and S_{worker} differ in the right part of the cue, where the method directly utilized $S_{\text{B\&S}}$ as the output. This difference occurs in the *because*-clause: S_{worker} included the clause as the scope but $S_{\text{converted}}$ did not. During the analysis, we found that whether or not a *because*-clause is included in the scope depends on the annotator's judgment, even following the guidelines of Sherlock: According to the principle that the scope of the cue ***not*** that negates a main verb includes subordinate clauses, the *because*-clause should be included in the scope. However, if an annotator applies *it is not the case that* paraphrase test[14] provided in the guidelines, which can be applied when doubts arise in annotation, the *because*-clause can be excluded from the scope. In fact, Sherlock contains instances with and without the *because*-clause annotated as part of the scope. The error pattern regarding clauses or phrases that express reason was frequently observed, suggesting that the relatively low recall scores in the token-level metrics (*hypo-token* and *scope-token*), as shown in Section 4.3, were partly due to these variations in scope annotation. Handling these annotation variations is difficult when designing a conversion method.

5 Negation Scope Resolution with a Unified Dataset

We conducted experiments on negation scope resolution to evaluate the effectiveness of the proposed method in terms of scaling up the dataset through our scope conversion. By adapting our conversion method to B&S, we obtained a converted version of B&S, namely, B&S whose scope annotation is almost identical to that of Sherlock. We merged the training sets of Sherlock and the converted version of B&S into a unified dataset. Using the unified dataset, we fine-tuned PLM-based models and performed negation scope resolution.

[14] For the detail of the *it is not the case that* paraphrase test, see Morante et al. (2011).

5.1 Experimental Settings

We fine-tuned BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019)[15] for negation scope resolution using the code of Truong et al. (2022).[16] The evaluation was performed under the different configurations of fine-tuning:

(i) **Sherlock only**: using only the training set of Sherlock.

(ii) **Combination**: using the training sets of Sherlock and the unconverted (original) version of B&S.

(iii) **Resolution + Combination (R + C)**: using the training set of Sherlock along with the training sets of B&S whose all scope labels were predicted by a scope resolution method. We used two different scope resolution methods: Yoshida et al. (2023) with our modifications and NegBERT.

(iv) **Conversion + Combination (C + C)**: using the training sets of Sherlock and the converted version of B&S, that is, using the unified dataset created using our conversion method. For the negation scope resolution process in the conversion method, we used the scope resolution method by Yoshida et al. (2023) with the modifications explained in Section 3.2. This is because the performance of scope conversion with Yoshida et al. (2023) outperformed that of scope conversion with NegBERT, as discussed in Section 4.3.

For the validation and the evaluation of the models, we used the validation and test sets of Sherlock, respectively. For all the settings, we adopted the same hyperparameters as those used by Truong et al. (2022).

We adopted two evaluation metrics, *scope-token* and *scope-strict-match*, both of which are explained in Section 4.1. For both metrics, we calculated the precision, recall, and F_1 measures. We adopted these metrics because they are the major metrics that have been used in many previous studies since their proposal in *SEM2012 shared task (Morante and Blanco 2012). The use of these metrics enables a fair comparison between the results of this study and those of the previous studies. Note that there are two differences used in this experiment and *SEM2012 shared task: in this experiment, punctuations are also included in the evaluation and gold cues are given. These differences stem from the adjustments made to align with the evaluation metrics used in the previous studies on PLM-based negation scope resolution.

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

[16] We referred to `https://github.com/joey234/negation-focused-pretraining`, which includes the re-implementation of Khandelwal and Sawant (2020). Note that the method of fine-tuning BERT for negation scope resolution was proposed by Khandelwal and Sawant (2020).

5.2 Experimental Results

Table 10 shows the experimental results. The results demonstrate that the models trained on the simply combined dataset of the three corpora (Combination) performed worse than those trained on the Sherlock dataset alone (Sherlock only). In contrast, the performance of the models improved when using the unified dataset (C + C). These results are observed in both models (BERT and RoBERTa) and in both evaluation metrics (*scope-token* and *scope-strict-match*), supporting the effectiveness of our conversion method in terms of scaling up the dataset. In addition, the models trained on C + C outperformed those trained on R + C (Yoshida et al.

Configuration/Method	Model	*Scope-token*			*Scope-strict-match*		
		Pre.	Rec.	F_1	Pre.	Rec.	F_1
Sherlock only*	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
Yoshida et al. (2023) +Combination	BERT	93.32	91.33	92.30	98.40	73.69	84.25
NegBERT+Combination	BERT	**95.13**	89.27	92.10	**99.45**	72.83	84.08
Conversion+Combination (our method)	BERT	93.84	92.43	**93.13**	98.91	<u>74.21</u>	84.79
Sherlock only*	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
Yoshida et al. (2023) +Combination	RoBERTa	89.84	<u>92.85</u>	91.31	98.95	60.53	75.10
NegBERT+Combination	RoBERTa	93.01	90.39	91.68	<u>99.33</u>	59.92	74.75
Conversion+Combination (our method)	RoBERTa	91.47	92.10	91.76	99.08	60.53	75.14
Khandelwal and Sawant (2020)*	BERT	—	—	92.36	—	—	—
Truong et al. (2022) – Baseline *	RoBERTa	—	—	91.51	—	—	—
Truong et al. (2022) – CueNB	RoBERTa	—	—	91.24	—	—	—
Wu and Sun (2023)	BERT	<u>95.12</u>	90.57	92.77	—	—	<u>85.35</u>
Wu and Sun (2023)	RoBERTa	94.54	91.24	<u>92.85</u>	—	—	**87.10**
Yoshida et al. (2023)**	heuristics	89.32	**94.30**	91.74	98.94	**74.70**	85.13

Table 10 Results for negation scope resolution with the different configurations of fine-tuning. The scores in our experiments are an average of five runs with different random seeds. Previous studies used only Sherlock for fine-tuning. Values marked in bold and underlined represent the highest and the second-highest scores for each evaluation metric, respectively.

* BERT fine-tuned on Sherlock only and RoBERTa on Sherlock only 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.

(2023) + Combination/NegBERT + Combination) for both the models and both evaluation metrics. This result suggests that our strategy of utilizing the existing scope labels in B&S is effective for improving the performance of PLM-based negation scope resolution models based on the scaling up of the dataset.

With regard to the *scope-token* metric, the models trained on the unified dataset (C + C) achieved lower precision and higher recall than the baseline, resulting in a higher F_1 measure. The higher recall of C + C indicates that the models assigned more scope labels by trained on more variants of Sherlock, and produced fewer False Negative predictions. Since PLM-based methods tend to have lower recall scores in *scope-token* than syntax-based methods (such as Yoshida et al. (2023)'s method), improving the recall score is crucial for improving the performance of PLM-based models. Therefore, the higher recall of C + C suggests that training on the unified dataset can enhance the performance of negation scope resolution models based on PLMs.

For both metrics, BERT consistently outperformed RoBERTa. This trend is consistent with the previous studies, showing that BERT achieves better performance in terms of negation scope resolution. For F_1 measures in *scope-token*, BERT trained on the unified dataset (C + C) outperformed the state-of-the-art models: Wu and Sun (2023), who proposed the model considering boundary shift loss, and Truong et al. (2022), who performed additional pre-training with negation. This finding demonstrates that the models can achieve high performance with basic fine-tuning on the unified dataset without requiring additional pre-training or using alternative model structures, in terms of *scope-token*. For *scope-strict-match*, the state-of-the-art models outperformed the results of our experiments. However, there is a possibility for the models to achieve even higher performance by fine-tuned on our unified dataset, which we leave the exploration for future work.

5.3 Analysis

To further investigate the effectiveness of scaling up the dataset through scope conversion, we conducted an analysis on the validation set of the Sherlock dataset. We classified the instances into several categories based on the types of negation cues and calculated the model performance for each category in the two settings: Sherlock only and C + C. The types of negation cues used for classification are as follows:

- Single word cue: negation cue that consists of a single word, such as "**not**" and "**no**".
- Affixal cue: negation cue that is an affix, such as "**im**(possible)" and "**un**(known)".
- Contracted cue: negation cue that is contracted with a verb, such as "(do)**n't**".
- Multiword cue: negation cue that consists of multiple words, such as "**no longer**".

The PLM model used for the analysis was BERT, which outperformed RoBERTa in the experiments on the test set (Table 10).

Tables 11 and 12 show the results of the analysis for the metrics *scope-token* and *scope-strict-match*, respectively. In the *scope-strict-match* metric, some categories contain excessively few instances to perform a meaningful analysis; therefore, our analysis focuses on the results of *scope-token*. The results in Table 11 demonstrate that, in terms of F_1 measure, the performance in the categories affixal and multiword cues improved with Conversion + Combination (C + C) compared to Sherlock only (SH) (+3.85 and +5.23 point, respectively); the performance for single word cues remained nearly unchanged; the results for contracted cues showed a decrease

Cue type	#Tokens	Pre.			Rec.			F_1		
		SH	C + C	Gain	SH	C + C	Gain	SH	C + C	Gain
Single word	2511 (69.90%)	99.21	94.46	−4.75	86.27	90.20	+3.93	92.29	92.28	−0.01
Affixal	654 (18.21%)	78.19	83.43	+5.24	84.48	86.78	+2.30	81.22	85.07	+3.85
Contracted	342 (9.52%)	99.30	91.56	−7.74	100.00	100.00	±0.00	99.65	95.59	−4.06
Multiword	85 (2.37%)	96.43	96.77	+0.34	81.82	90.91	+9.09	88.52	93.75	+5.23
All	3592	95.98	92.69	−3.29	87.35	90.79	+3.44	91.47	91.73	+0.26

Table 11 Comparison of *scope-token* performance for each type of negation cue between Sherlock only (referred to as SH here) and Conversion + Combination (C + C).The base model is BERT, and performance was measured in the validation set.

Cue type	#Scopes	Pre.			Rec.			F_1		
		SH	C + C	Gain	SH	C + C	Gain	SH	C + C	Gain
Single word	116 (67.05%)	100.00	100.00	±0.00	75.68	79.28	+3.60	86.15	88.44	+2.29
Affixal	32 (18.50%)	100.00	100.00	±0.00	40.63	40.63	±0.00	57.78	57.78	±0.00
Contracted	20 (11.56%)	100.00	100.00	±0.00	95.00	95.00	±0.00	97.44	97.44	±0.00
Multiword	5 (2.89%)	100.00	100.00	±0.00	60.00	60.00	±0.00	75.00	75.00	±0.00
All	173	100.00	100.00	±0.00	70.83	73.21	+2.38	82.93	84.54	+1.61

Table 12 Comparison of *scope-strict-match* performance for each type of negation cue between Sherlock only (referred to as SH here) and Conversion + Combination (C + C). The base model is BERT, and performance was measured in the validation set.

in performance (−4.06 point). Consequently, along with the performance for Sherlock only, our method contributed to enhancing the performance in the categories where Sherlock only had relatively low performance. We observed the most significant performance gains for affixal cues. Since there are various kinds of affixal cues, it is considered that our dataset scaling up leads to performance improvement in this category. However, the results for contracted cues showed decreased performance in C + C compared with Sherlock only. A possible reason for this decrease is the relatively poor performance of our conversion method for the scopes of contracted cues. This is a limitation of the proposed method in terms of scaling up the dataset, which should be addressed in future work.

6 Conclusion

This study proposed a method for automatically converting the negation scopes of B&S to those of Sherlock, utilizing the existing scope annotations in B&S. We conducted an experiment to evaluate the accuracy of the proposed method. The results demonstrated that our method can convert the scopes of B&S to those of Sherlock with high accuracy and that our approach of utilizing the correct scopes of B&S is effective. Using the proposed method, we merged B&S and Sherlock into a unified dataset. To verify the effectiveness of our method in terms of scaling up the dataset, we conducted experiments of negation scope resolution using the unified dataset for fine-tuning PLM-based models. The experimental results demonstrated that the unified dataset improves the performances of the models, unlike a simple combination of the corpora. Potential future works include exploring the reverse scope conversion: converting the scopes of Sherlock to those of B&S. By automatically identifying and excluding the left parts of the negation scope to the cue(s) in Sherlock that are not annotated in B&S, we expect to convert the scopes of Sherlock to those of B&S. Examining the details of this method will be a topic for future research. Another future work involves 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.

Acknowledgement

This research was partially supported by the Grant-in-Aid for Scientific Research (C) (No. 22K12148) of JSPS.

References

Bird, S. and Loper, E. (2004). "NLTK: The Natural Language Toolkit." In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pp. 214–217, Barcelona, Spain. Association for Computational Linguistics.

Chen, J., Shi, W., Fu, Z., Cheng, S., Li, L., and Xiao, Y. (2023). "Say What You Mean! Large Language Models Speak Too Positively about Negative Commonsense Knowledge." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 9890–9908, Toronto, Canada. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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*, pp. 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fancellu, F., Lopez, A., Webber, B., and He, H. (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*, pp. 58–63, Valencia, Spain. Association for Computational Linguistics.

García-Ferrero, I., Altuna, B., Alvez, J., Gonzalez-Dios, I., and Rigau, G. (2023). "This is not a Dataset: A Large Negation Benchmark to Challenge Large Language Models." In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8596–8615, Singapore. Association for Computational Linguistics.

Hossain, M. M., Chinnappa, D., and Blanco, E. (2022). "An Analysis of Negation in Natural Language Understanding Corpora." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 716–723, Dublin, Ireland. Association for Computational Linguistics.

Khandelwal, A. and Sawant, S. (2020). "NegBERT: A Transfer Learning Approach for Negation Detection and Scope Resolution." In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 5739–5748, Marseille, France. European Language Resources Association.

Kitaev, N., Cao, S., and Klein, D. (2019). "Multilingual Constituency Parsing with Self-Attention and Pre-Training." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3499–3505, Florence, Italy. Association for Computational Linguistics.

Kitaev, N. and Klein, D. (2018). "Constituency Parsing with a Self-Attentive Encoder." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Konstantinova, N., de Sousa, S. C., Cruz, N. P., Maña, M. J., Taboada, M., and Mitkov, R. (2012). "A Review Corpus Annotated for Negation, Speculation and Their Scope." In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pp. 3190–3195, Istanbul, Turkey. European Language Resources Association.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach." *arXiv preprint arXiv:1907.11692*.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). "Building a Large Annotated Corpus of English: The Penn Treebank." *Computational Linguistics*, **19** (2), pp. 313–330.

Morante, R. and Blanco, E. (2012). "*SEM 2012 Shared Task: Resolving the Scope and Focus of Negation." In **SEM 2012: The 1st Joint Conference on Lexical and Computational Semantics*, pp. 265–274, Montréal, Canada. Association for Computational Linguistics.

Morante, R. and Daelemans, W. (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*, pp. 1563–1568, Istanbul, Turkey. European Language Resources Association.

Morante, R., Schrauwen, S., and Daelemans, W. (2011). "Annotation of Negation Cues and Their Scope: Guidelines v1.0." Tech. rep., University of Antwerp.

Ravichander, A., Gardner, M., and Marasovic, A. (2022). "CONDAQA: A Contrastive Reading Comprehension Dataset for Reasoning about Negation." In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8729–8755, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Read, J., Velldal, E., Øvrelid, L., and Oepen, S. (2012). "UiO1: Constituent-Based Discriminative Ranking for Negation Resolution." In **SEM 2012: The 1st Joint Conference on Lexical and Computational Semantics*, pp. 310–318, Montréal, Canada. Association for Computational Linguistics.

She, J. S., Potts, C., Bowman, S. R., and Geiger, A. (2023). "ScoNe: Benchmarking Negation Reasoning in Language Models With Fine-Tuning and In-Context Learning." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 1803–1821, Toronto, Canada. Association for Computational Linguistics.

Szarvas, G., Vincze, V., Farkas, R., and Csirik, J. (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*, pp. 38–45, Columbus, Ohio. Association for Computational Linguistics.

Truong, T., Baldwin, T., Cohn, T., and Verspoor, K. (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*, pp. 4188–4193, Seattle, United States. Association for Computational Linguistics.

Truong, T. H., Baldwin, T., Verspoor, K., and Cohn, T. (2023). "Language Models are not Naysayers: An Analysis of Language Models on Negation Benchmarks." In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics*, pp. 101–114, Toronto, Canada. Association for Computational Linguistics.

Weller, O., Lawrie, D., and Van Durme, B. (2024). "NevIR: Negation in Neural Information Retrieval." In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2274–2287, St. Julian's, Malta. Association for Computational Linguistics.

Wu, Y. and Sun, A. (2023). "Negation Scope Refinement via Boundary Shift Loss." In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6090–6099, Toronto, Canada. Association for Computational Linguistics.

Ye, M., Kuribayashi, T., Suzuki, J., Kobayashi, G., and Funayama, H. (2023). "Assessing Step-by-Step Reasoning against Lexical Negation: A Case Study on Syllogism." In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14753–14773, Singapore. Association for Computational Linguistics.

Yoshida, A., Kato, Y., and Matsubara, S. (2023). "Revisiting Syntax-Based Approach in Negation Scope Resolution." In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics*, pp. 18–23, Toronto, Canada. Association for Computational Linguistics.

Yoshida, A., Kato, Y., and Matsubara, S. (2024). "Negation Scope Conversion: Towards a Unified Negation-Annotated Dataset." In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pp. 12093–12099, Torino, Italia. ELRA and ICCL.

Appendix

A Application Frequency of Each Path Pattern Rules

Table 13 shows the application frequency of each path pattern rule in the training sets of BioScope and SFU.

Path pattern rule	BioScope		SFU	
RB//VP/S/SBAR if SBAR\WHNP	39	(2.66%)	91	(3.68%)
RB//VP/S	410	(27.95%)	1246	(50.38%)
RB//S	231	(15.75%)	349	(14.11%)
DT/NP if NP/PP	3	(0.20%)	24	(0.97%)
DT//SBAR if SBAR\WHADVP	2	(0.14%)	6	(0.24%)
DT//S	182	(12.41%)	270	(10.92%)
JJ//ADJP/VP/S if S\VP\VB* [@lemma="be"]	11	(0.75%)	0	(0.00%)
JJ/NP/NP if NP\PP	0	(0.00%)	0	(0.00%)
JJ//NP	3	(0.20%)	4	(0.16%)
UH	0	(0.00%)	0	(0.00%)
IN/PP	80	(5.45%)	105	(4.25%)
NN//NP/NP if lemma of NN in ["lack", "absence"]	68	(4.64%)	1	(0.04%)
NN/NP//S/SBAR if SBAR\WHNP	1	(0.07%)	1	(0.04%)
NN/NP//S	20	(1.36%)	39	(1.58%)
CC/SINV	2	(0.14%)	2	(0.08%)
VBG//NP	20	(1.36%)	0	(0.00%)
VBN//NP	0	(0.00%)	2	(0.08%)
VB*//S	68	(4.64%)	75	(3.03%)
No rule activated	327	(22.29%)	258	(10.43%)
All	1467		2473	

Table 13 Application frequency of each path pattern rule (Figure 5) in the training sets of BioScope and SFU. The underlined rules are those that we added in this study. The count was measured per instance (i.e., per negation cue).

Asahi Yoshida: Asahi Yoshida received his bachelor's degree in informatics from Nagoya University in 2023. He is currently a master student at Graduate School of Informatics, Nagoya University. His research interests include natural language processing, especially neagtion processing in natural language texts.

Yoshihide Kato: Yoshihide Kato received the B.E. degree, the M.E. degree, and the Dr. of Engineering from Nagoya University in 1997, 1999, and 2003, respectively. He is currently an Associate Professor at Information & Communications, Nagoya University. His research interests include natural language processing and computational linguistics.

Shigeki Matsubara: Shigeki Matsubara received the B.E. degree from Nagoya Institute of Technology in 1993, and the M.E. degree, and the Dr. of Engineering from Nagoya University in 1995 and 1998, respectively. After becoming an Assistant Professor and an Associate Professor at Nagoya University, he became a Full Professor in 2017. His research interests include natural language processing and digital library.

(Received August 1, 2024)
(Revised November 11, 2024)
(Accepted December 23, 2024)