Solving Label Variation in Scientific Information Extraction via Multi-Task Learning

Dong Pham[1], **Xanh Ho**[2], **Quang-Thuy Ha**[1] and **Akiko Aizawa**[2,3]
[1]VNU University of Engineering and Technology, Hanoi, Vietnam
[2]National Institute of Informatics, Tokyo, Japan
[3]The University of Tokyo, Tokyo, Japan
dongpham120899@gmail.com
thuyhq@vnu.edu.vn
{xanh, aizawa}@nii.ac.jp

Abstract

Scientific Information Extraction (ScientificIE) is a critical task that involves the identification of scientific entities and their relationships. The complexity of this task is compounded by the necessity for domain-specific knowledge and the limited availability of annotated data. Two of the most popular datasets for ScientificIE are SemEval-2018 Task-7 and SciERC. They have overlapping samples and differ in their annotation schemes, which leads to conflicts. In this study, we first introduced a novel approach based on multi-task learning to address label variations. We then proposed a soft labeling technique that converts inconsistent labels into probabilistic distributions. The experimental results demonstrated that the proposed method can enhance the model robustness to label noise and improve the end-to-end performance in both ScientificIE tasks. The analysis revealed that label variations can be particularly effective in handling ambiguous instances. Furthermore, the richness of the information captured by label variations can potentially reduce data size requirements. The findings highlight the importance of releasing variation labels and promote future research on other tasks in other domains. Overall, this study demonstrates the effectiveness of multi-task learning and the potential of label variations to enhance the performance of ScientificIE.[1]

1 Introduction

Information extraction (IE) refers to the process of automatically identifying the entities and relations from unstructured text. Extracting the information from a scientific paper is more challenging than from open-domain data, given that scientific texts require in-depth knowledge of the subject matter for accuracy; thus, labeling is costly and the amount of labeled data is limited. Bassignana and Plank (2022b) revealed that two well-known datasets, namely, SemEval-2018 (Gábor et al., 2018) and SciERC (Luan et al., 2018), contain overlapped abstracts and directly correspondent labels; however, they differ in their annotations. In particular, there are 307 abstracts (out of 500 abstracts) that are overlapped between the two datasets. The number of annotated relations in these abstracts differs significantly and includes conflicting instances. The presence of conflicting annotations raises concerns regarding the reliability of these datasets; thus, the determination of trustworthiness is challenging, especially with limited resources.

Labeling plays a crucial role in machine learning pipelines, as it involves assigning labels to data points to train models effectively. However, human labeling is generally subject to errors and inconsistencies (Plank, 2022), and label aggregation cannot capture the actual complexity of the world (Basile et al., 2021). Using only high-agreement instances for model training and testing can cause overfitting and data redundancy (Jamison and Gurevych, 2015). Therefore, different annotated opinions should be retained and "variation" should be considered over "disagreement", given that disagreement annotations imply that two (or more) views involved are not all accurate (Plank, 2022).

Label variation occurs in ScientificIE when different annotators assign different labels to the same entity or relationship. This variation can stem from various factors, including differences in domain knowledge or interpretations of annotation guidelines, in addition to the subjective understanding of the underlying data. In response to the challenge of label variations arising from overlapping datasets, we developed a novel approach based on multi-task learning. By jointly training the proposed model on multiple perspectives, overlapping and conflicting annotations can be effectively handled. We released soft labels (a probability distribution generated by multi-level agreements) as an auxiliary

[1]Data and code are publicly available at: https://github.com/dongpham120899/LabelVariation_SciIE

loss. Leveraging soft labels with several loss functions can reduce the penalty for errors and enhance the model robustness (Fornaciari et al., 2021).

To evaluate the effectiveness of the proposed approach, we conducted experiments using overlapped data as the training set and non-overlapped data as the testing set. We compared the performances of models trained on these datasets using traditional label aggregation methods and the proposed multi-task learning approach. Additionally, we conducted a cross-dataset evaluation on the SciREX dataset (Jain et al., 2020) and performed testing using the standard splitting of the SciERC benchmark (Luan et al., 2018). The experimental results revealed that the proposed approach effectively mitigated the impact of label variation on model performance, thus leading to improvements in the accuracy and robustness of two of the tasks, namely, name entity recognition (NER) and relation extraction (RE). In particular, we found that label variation is particularly effective in handling ambiguous instances, and the richness of information captured by label variation can reduce data size requirements.

Overall, the findings suggest that multi-task learning and soft labels derived from inconsistent annotations can be powerful tools for addressing label variations in ScientificIE tasks. Moreover, future research in this field should be promoted, to comprehensively consider the potential benefits of these approaches.

2 Label Variation in ScientificIE

Inconsistencies in annotations across different datasets can result in label variation, which presents a significant challenge for accurate and reliable machine learning models. The issue of label variation is exemplified in the overlap and annotation divergence observed in the SemEval-2018 Task 7 (Gábor et al., 2018) and SciERC (Luan et al., 2018) datasets, which has been discussed in previous research (Bassignana and Plank, 2022b). In this section, we delve deeper into the issue of inconsistent labels and argue for the importance of releasing variation labels in ScientificIE.

2.1 Datasets

SemEval-2018 Task 7 (Gábor et al., 2018) This dataset[2] comprises 500 abstracts from published research papers from the ACL Anthology. It focuses on predicting relations between two entities with six pre-defined relations (*Usage, Result, Model, Part-Whole, Topic, Comparison*). The entity annotations are first automatically identified and then manually corrected by other annotators. The target is to identify maximum noun phrases, abbreviations, and their context. The relation annotation process is divided into three steps: defining, validation, and annotation. The domain experts only annotate the semantic relations that are explicit and relevant to comprehending the abstract.

	SemEval-2018	**SciERC**
Label mapping	Comparison Usage Part-whole Model Result	Compare Used-for Part-of Feature-of Evaluate-for
Statistic on whole corpus		
# Entities	7483	8089
# Relations	1583	4648
# Relations/Doc	3.2	9.3
Statistic on 307 overlapped abstracts		
# Entities	4592	4252
# Relations	1087	2476
# Common Relations	1071	1922

Table 1: Label mapping between the two datasets and statistics in both datasets.

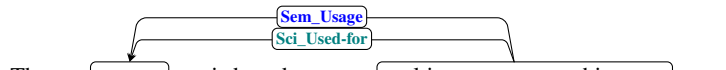
SciERC (Luan et al., 2018) This corpus includes annotations for scientific entities, their relations, and coreference for 500 scientific abstracts from the AI communities. They defined six types for scientific annotation entities (*Method, Metric, Task, Material, Generic, OtherScientificTerm*) and seven relation types (*Used-for, Evaluate-for, Feature-of, Part-of, Compare, Hyponym-of, Conjunction*). The final annotations were obtained by greedy strategy from multiple annotators. Their annotators were preferred to indicate a longer span whenever ambiguity occurs and ignore negative relations.

2.2 Overlap of the Datasets

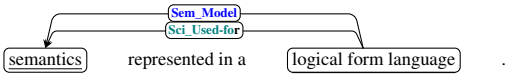
Bassignana and Plank (2022b) identified 307 abstracts that were common to both the SemEval-2018 Task 7 and SciERC datasets. This indicates that there are 193 non-overlapped abstracts in each dataset. In addition, most of the relationships in both datasets have direct corresponding labels. To clarify the correspondence, we computed the co-

[2]To ensure a fair comparison with SciERC, we utilized resources specific to sub-task 2 (Relation extraction and classification on clean data) from SemEval-2018 Task 7.

Example 1: Overlapped relation

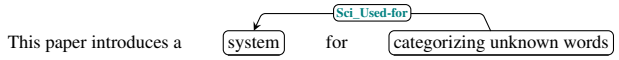


Example 2: Conflicted relation



Example 3: Conflicted entity

SemEval-2018: This paper introduces a system for categorizing unknown words .



Example 4: Different entity and relation

SemEval-2018: We propose a detection method . for orthographic variants caused by transliteration . in a large corpus .

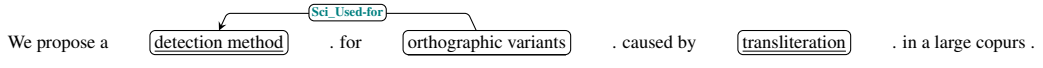


Table 2: The noise samples occur in label variation. For the relation part, we use "Sem" [*blue] to denote relation in the SemEval dataset, while "Sci" [*green] denotes the SciERC dataset.

occurrence score of relational labels between two pairs of entity labels, which is detailed in the Appendix A.

Table 1 provides an overview of the label mapping and the quantities of entities and relations in each dataset. Both datasets contain an equal number of abstracts; however, there is a minor disparity in the number of entities. The significant distinction arises from the amount of annotated relations, i.e., SemEval-2018 has 3.2 relations per abstract while SciERC has 9.3 relations per abstract. We further highlighted their distinctiveness in the distribution of common relations, as shown in Figure 3 in Appendix B. This inconsistency can be attributed to the different interpretations of annotation guidelines, where SemEval-2018 is focused on explicit relationships while SciERC on broader coverage

2.3 Release the Label Variation

The presence of label variation introduces inconsistency and ambiguity into the labeled data, which poses significant challenges for the training of accurate and reliable scientific extraction systems. Table 2 presents four noise scenarios between the two datasets. These examples illustrate the difficulties associated with resolving significant disagreements and the limitations of dependence on the gold label. The actual world is excessively complex to be represented by an independent perspective. Nonetheless, incorporating labels from both datasets to train a single model poses a significant challenge. Section 3 presents our proposed method that leverages multi-task learning to effectively address label inconsistencies arising from dataset overlaps.

3 Multi-Task Learning to Handle with Label Variation

In this section, we first present a summary of the architecture of the end-to-end model for IE. We then introduce our proposed method for the multi-task learning of multi-perspectives. Finally, we developed the soft label with multi-level agreements from inconsistent annotations to enhance the model's robustness.

3.1 SpERT

Eberts and Ulges (2019) introduced a span-based joint entity and relation extraction model referred to as SpERT, which is built upon the transformer pre-training framework. The authors highlighted the significance of localized context representation between entity pairs, which contributes to the effectiveness of their model. Furthermore, SpERT efficiently extracted a sufficient number of strong negative samples in a single BERT (Devlin et al., 2019) pass during training. Finally, SpERT out-

performed previous approaches on several datasets for the joint entity and relation extraction tasks. By employing joint modeling, SpERT effectively captured dependencies between entities and their relations, thus resulting in improved performance and reduced processing time.

3.2 Learning with Multi-Perspectives

We propose an approach that utilizes two output heads in the SpERT architecture, i.e., two in NER and two in RE, where each represents a single perspective in variation annotation. This extension allows our model to address challenges such as overlapping and conflicting instances within the same input text, enabling it to learn inconsistencies in an end-to-end manner for both NER and RE tasks. To achieve this multi-perspective learning, we introduce a unified loss function that jointly optimizes entity classification and relation classification. The joint loss function is expressed as follows:

$$L_i = L_i^s + L_i^r \quad (1)$$

$$L_{multi} = L_1 + L_2 \quad (2)$$

where L_i denotes the main loss trained with the i-th perspective annotation, L^s denotes the span classifier loss (cross-entropy over the entity classes including none), and L^r denotes the binary cross-entropy over the relation classes.[3] The multi-perspective loss is calculated by the sum of the single perspective loss, which encompasses both the NER and RE. To put it briefly, we used SpERT as the backbone to compute multi-perspectives, illustrated in Figure 1.

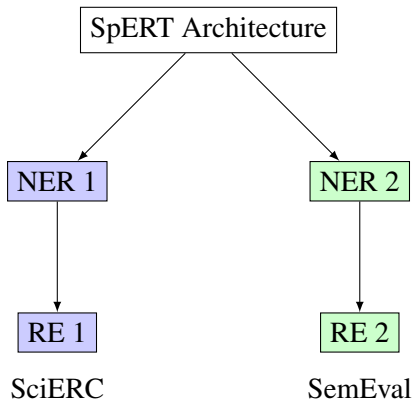


Figure 1: Illustration of Multi-Task Learning based on SpERT architecture to handle label variation.

[3]In this context, the loss function used by SciERC is denoted by L_1, while the loss function used by SemEval is denoted by L_2.

3.3 Soft Label from Multi-level Agreements

Unlike most existing models that rely on one-hot encoded gold distributions, Fornaciari et al. (2021) was based on a different approach, in that probability distributions were collected over the labels provided by annotators. This allowed for a more nuanced notion of truth by comparing it with soft labels. In the overlapping examples between SemEval-2018 and SciERC, we observed both consistent and inconsistent relations, which were considered as multi-level agreements. Example 1 in Table 2 demonstrates a high level agreement in annotations between the two datasets, whereas Example 2 illustrates conflicting relations with a low level of agreement. Several instances exhibited no similarities in entity pairs, thus resulting in different relations, as shown in Example 3. We utilized soft labels as probability distributions over the labels provided by the multi-level agreements to address these variations.

In this study, we introduced soft labels at three levels of agreement (high, medium, and low). The soft labels were manually computed based on the degree of agreement between the two sets of data. To provide a clearer illustration, consider the label "Sci_Used-for" in the first three examples in Table 2. In the first example, with high agreement, the label was assigned soft labels as distributions [0.9, 0.025, 0.025, 0.025, 0.025]. In the second example, with low agreement, the label was assigned [0.6, 0.1, 0.1, 0.1, 0.1]. Lastly, in the third example, with the medium agreement, the label was assigned probability distributions [0.8, 0.05, 0.05, 0.05, 0.05].[4]

To measure the difference between the predicted distribution Q and distribution of soft labels P, we used the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951). The standard KL-divergence is expressed as follows:

$$D_{KL}(P||Q) = \sum_x P(x) \log(\frac{P(x)}{Q(x)}) \quad (3)$$

The D_{KL} describes the amount of information lost when the distribution Q is used to approximate distribution P. Moreover, the inverse KL-divergence (Fornaciari et al., 2021) was introduced to encourage a narrow Q distribution, which causes

[4]In this case, we only consider five types of relationship labels, with the encoding order being [Use-for, Compare, Feature-of, Part-of, Evaluate-for].

the model to learn a distribution that directs attention toward the classes wherein annotations exhibit potential agreement. We also attempted to compute the soft label with cross-entropy or binary cross-entropy loss, as experimentally demonstrated.

We incorporated a soft label as an auxiliary task to mitigate overfitting. Each individual perspective was assigned its own soft label, and we included two auxiliary losses in addition to the multi-perspective loss. By considering the soft label between the two perspectives, closer alignment and facilitate learning were achieved. The final loss can be expressed as follows:

$$L_{soft} = D_{KL}(P_1||Q_1) + D_{KL}(P_2||Q_2) \quad (4)$$

$$L_{multi\_with\_soft} = L_{multi} + L_{soft} \quad (5)$$

To avoid underflow issues during computation, we applied logarithmic normalization to the soft label. Additionally, we utilized the LogSoftmax activation function for the auxiliary loss, thus ensuring that the probabilities of the individual labels did not approach zero.

4 Experiments

To assess the effectiveness of the proposed method, we applied three main experimental scenarios to three datasets. In the first setup, we utilized the overlaps between two datasets as a training set and evaluated two of the non-overlaps as testing sets in the RE task. We then performed a cross-dataset evaluation on the SciREX dataset (Jain et al., 2020) for the NER task. Finally, we evaluated the proposed method on the SciERC leaderboard (Luan et al., 2018) in both tasks.

For all the experiments, we leveraged the SciBERT (cased) model (Beltagy et al., 2019) as a sentence encoder, i.e., a BERT model pre-trained on a large corpus of scientific papers. We used the SpERT architecture as a baseline model to run other training sets with the same hyper-parameters reported in Eberts and Ulges (2019). We utilized the spaCy toolkit (Honnibal and Montani, 2017) to split abstracts into sentences because both SemEval and SciERC datasets only have relations within sentences, and SpERT requires a single sentence as input. It is noted that the reported score is the average score from five runs that use different seeds.

4.1 Overlap and Non-overlap

4.1.1 Deal with Label Variation

Two datasets with two annotation perspectives lead to inconsistencies in ScientificIE. In this experimental setup, we utilized 307 overlapped abstracts with multiple annotation perspectives to train the model, and the two non-overlapped sets in SemEval and SciERC were used as two testing sets. However, it is important to note that SemEval-2018 did not provide entity types in their dataset. Consequently, we removed all entity types from SciERC and focused only on the RE task with five common relationships, as shown in Table 1. A performance evaluation was conducted in both sets using the micro F1-score metric, and the final score was obtained by averaging the results.

By leveraging the corresponding labels, our primary objective was to investigate the adaptive capabilities of the two datasets through cross-evaluation (1.1 and 1.2). With the overlap, we attempted to implement other means to handle label variations in accordance with training configurations:

- *concat*: We employed concatenation by including each abstract twice: once from each dataset (2.1). This technique allowed for the incorporation of sentence inputs with different annotations, thus providing a simple method to address variations in annotation (Sheng et al., 2008; Uma et al., 2021b). Additionally, to increase the amount of training data, we leveraged the non-overlapping exclusions in the respective testing sets (2.2 and 2.3).

- *mix*: We doubled the annotation of the abstracts from two datasets. In the presence of overlapped relations, the consistent relation was retained. With the conflicting annotations, we filtered out the SemEval-2018 relation in (3.2) and the SciERC relation in (3.3).

Results Table 3 reports the micro F1-scores obtained from the testing sets, including SemEval, SciERC, and their average. Training on independent datasets, (1.1) and (1.2) yielded satisfactory results on only one of the sets, indicating limited adaptability. The concatenation approach (2.1) to increase training data led to decreased performance due to inconsistency in the data. Further testing with non-overlapping portions of the datasets, (2.2) and (2.3) resulted in a further drop in performance. With respect to the mixed labeling approach, we

No.	Test set	SemEval-2018			SciERC			Average		
	Train set	*Precision*	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*	*Precision*	Recall	F1-score
1.1	SemEval	21.39	22.93	22.13	27.30	9.33	13.91	24.35	16.13	18.02
1.2	SciERC	6.24	13.84	8.60	43.02	35.83	39.10	24.63	24.84	23.85
2.1	Concat set	11.73	23.35	15.62	41.87	27.99	33.55	26.80	25.67	24.59
2.2	Concat set + Sci	10.22	23.55	14.25	-	-	-	-	-	-
2.3	Concat set + Sem	-	-	-	39.80	27.36	32.43	-	-	-
3.1	Mixed set	9.76	**29.13**	14.62	35.14	35.20	35.17	22.45	32.17	24.89
3.2	Mixed-Sci set	9.44	28.10	14.69	36.35	**36.35**	36.35	22.90	**32.23**	25.52
3.3	Mixed-Sem set	10.10	28.72	14.95	34.99	33.68	33.80	22.54	31.20	24.28
4.1	*MTL	23.97	19.90	21.74	**44.62**	34.00	38.60	34.27	26.95	30.17
4.2	*MTL with soft label	**24.69**	20.75	**22.37**	44.33	35.46	**39.66**	**34.51**	28.11	**31.02**

Table 3: Micro F1-scores of the experimental training on the overlap data and testing on the non-overlap data. 1.1 and 1.2 refer to independent training data. 2.1, 2.2, and 2.3 represent the cases where we repeat abstracts from two datasets. 3.1, 3.2, and 3.3 indicate double annotation. 4.1 and 4.2 represent our proposed model.

found that the mixed dataset achieved the highest recall score. Combining the two types of annotations increased the number of predictions, thus resulting in a small number of false negatives. This led to an increase in the recall score and a decrease in the precision score, which impacted the overall F1-score. Additionally, by removing conflicting relations from the mixed dataset, such as prioritizing either SciERC or SemEval annotations (3.2) or (3.3), we observed a slight improvement in performance. This emphasized the inherent limitation of traditional models in effectively capturing and incorporating multiple perspectives. In contrast, the proposed method that utilizes multi-task learning with soft labels, achieved the optimal F1-score and precision score. It should be noted that when using multi-task learning without soft labels (4.1), the performance in individual testing sets was not superior to that when training on independent sets (Sem: 21.74 vs. 22.13 and Sci: 38.60 vs. 39.10). However, the incorporation of soft labels (4.2) improved the performance, particularly in addressing label variations and enhancing the model robustness to inconsistencies (Sem: 21.74 $\rightarrow$ 22.37 and Sci: 38.60 $\rightarrow$ 39.66). When comparing the average scores, the multi-task learning approach outperformed other methods in handling label variations within overlapped datasets.

4.1.2 Impact of Data Quantity

The success of training a deep learning model relies significantly on the availability of sufficient training data. In the context of ScientificIE, the limited availability of data can be attributed to the requirement of expert labeling. Obtaining a sizeable amount of such variations in label data is increasingly challenging (1400 sentences in 307 overlapped abstracts). In a previous study (Zhang et al., 2021), new annotation distribution schemes were investigated with respect to the learning of multiple labels per example for a small subset of training examples, which can lead to novel architectures. Moreover, Plank (2022) revealed the potential of label variations to reduce data size. Thus, we conducted a comparison within the SciERC testing set between the gold and variation labels when decreasing data quantity.

Results Figure 2 illustrates the downward trend in performance in accordance with a decrease in the quantity of data in intervals of 100 training samples. The performance of the gold labels (shown in red) exhibited a rapid decrease when the data amount was reduced from 1400 to 1100. In contrast, the variation labels trained using the proposed method (shown in blue) exhibited minimal changes within the same range. Both methods exhibited a similar decline in performance when the dataset size was excessively small (from 400–1000 samples). This observation demonstrated that the richness of

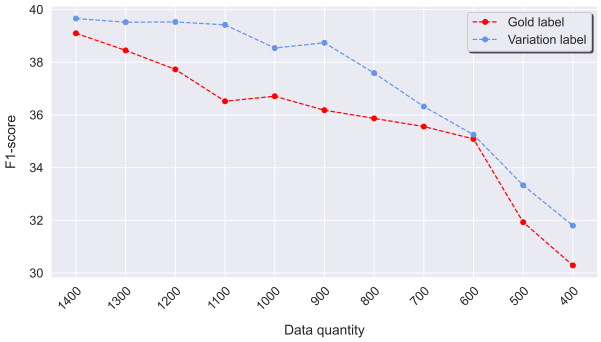


Figure 2: The impact of data quantity on the performance (SciERC testing). The gold label was trained on SciERC annotations with the SpERT baseline model, and the variation label was trained using the proposed model.

Model	RE		
	Precision	*Recall*	*F1-score*
MTL	44.62	34.00	38.60
MTL + BCE	41.94	34.74	38.01
MTL + CrossEntropy	42.65	34.86	38.36
MTL + KL-inverse	**45.76**	34.97	39.65
MTL + KL-standard	44.33	**35.46**	**39.66**

Table 4: The performance of multi-task learning with (MTL row) and without soft labels (the rest). We compared the effects of different loss functions for soft labels, including the following: BCE, cross-entropy, KL-inverse (Fornaciari et al., 2021), and KL-standard. All the scores in this table are obtained by evaluating the model on the SciERC testing set.

information captured by label variations remained stable, even with smaller datasets. Achieving a tradeoff between the data quantity and label diversity is a critical consideration for maximizing the effectiveness of machine learning models in various tasks.

4.1.3 Impact of Soft Label

We presented the soft labeling developed from multi-level agreements, as shown in Section 3.3, and used other loss functions to capture the soft loss. Fornaciari et al. (2021) proposed an inverse version of the KL-divergence, which was unsuccessful when we applied the logarithmic norm; thus, we computed the inverse version without normalization. Moreover, we attempted a standard cross-entropy with a softmax activation function at the final output head and a standard Binary-Cross-Entropy (BCE) with a sigmoid layer at the output head.

Results Table 4 reports the micro-F1 scores on the SciERC testing set of the MLT model with and without soft label. Using BCE and cross-entropy to measure the difference between the prediction distribution and the distribution of soft labels reduced the performance when compared with the case wherein soft labels were not used (only MTL). In this study, the inverse version of KL-divergence was not superior to the standard version with logarithmic normalization. There were few differences between the two versions. Overall, KL led to consistent performance improvements in soft labels.

4.2 Cross-Dataset Evaluation on SciREX

SciREX (Jain et al., 2020) is a document-level IE dataset for scientific articles, which covers tasks such as entity identification and N-ary relation extraction. In particular, it combines automatic and human annotations, thus leveraging existing scientific knowledge resources.

In the Table 1, there are still differences in entity annotations between the two datasets. SciERC annotators indicate the long span, including prepositions. While entity annotations in SemEval-2018 are indicated maximal noun phrases, abbreviations, etc., they often are shorter (Example 3 or 4 in Table 2). To evaluate the cross-dataset, we selected 368 abstracts from full-text papers in the SciREX dataset and investigated the NER task on four entity types (Method, Task, Metric, and Material) released by the SciREX dataset. We only used the abstract section, given that the SemEval or SciERC dataset only uses the abstract. We trained the SpERT model on gold labels with SciERC annotations. With variations, we retained the entity annotations of SciERC (six entity types) and released a new entity type denoted as "OtherScientificTerm_2"[5] for SemEval-2018 annotations. Entity prediction results were obtatined at the head of the SciERC prediction, and the final scores were calculated by micro-averaging four entity types.

Results We compared the performances between the gold labels and variation labels of SciERC annotations, and the results revealed in Table 5. The two models trained on label variations outperformed the model trained on the gold labels (>1%). This

[5]SemEval-2018 didn't release entity types in their dataset.

Model	Train set	NER *F1-score*
SpERT	SciERC's gold label	43.31
SpERT_MTL	Variation label	44.37
SpERT_MTL + soft label	Variation label	**44.64**

Table 5: Micro F1-score of cross-dataset evaluation in the NER task.

highlights the effectiveness of combining two types of annotations from the overlap, as it leads to improved performance compared with using a single annotation. Additionally, we observed the impact of using soft labels in this experiment, thus further emphasizing their effectiveness in addressing label inconsistencies.

Model	Label	NER *F1-score*	RE *F1-score*
Pipeline model			
PL.Marker	gold	69.90	**53.20**
End-to-End model			
SpERT_MTL + soft label	variation	**70.83**	51.31
SpERT.PL	gold	70.53	51.25
SpERT_MTL	variation	70.61	51.02
SpERT	gold	70.30	50.84

Table 6: The comparison of existing methods on the leaderboard of SciERC, underlined is our method.

4.3 Standard Splitting SciERC

The SciERC benchmark (Luan et al., 2018) consists of two sets: a training set and a testing set. Among 307 overlapped abstracts, 252 abstracts are included in the training set (400 abstracts) and 55 in the testing set (100 abstracts). Due to this overlap, we were unable to obtain variation labels for the entire training dataset, which can be considered as a disadvantage. Using the proposed method, we trained both the overlapping and non-overlapping samples using two tasks and two types of annotations. Among the 148 non-overlapping abstracts, we retained those with medium-level agreement in the soft labels. The experimental setup was identical to the one described in section 4.2. The micro F1-scores for both tasks were calculated based on the gold label annotations from SciERC.[6]

Results The performance of the proposed method on the SciERC leaderboard is presented in Table 6. The proposed approach surpassed the state-of-the-art models in entity recognition, thus achieving an improvement of 0.6–0.8% when compared with the SpERT (Eberts and Ulges, 2019) model and its variant, i.e., SpERT.PL (Santosh et al., 2021). The incorporation of diverse entity annotations, even with minor conflicts, is beneficial for enhancing the accuracy of the NER task. In the RE task, the proposed method achieved the highest F1-score among existing end-to-end models. However, the improvements were limited due to significant conflicts in the relation annotations between the two datasets and the testing set as the gold label of SciERC. Furthermore, the proposed approach did not outperform a pipeline model (Ye et al., 2022) in the RE task. Overall, the proposed method, which utilizes multi-task learning with soft labels based on the SpERT architecture, enhanced the performance of the baseline model in both tasks.

5 Error Analysis

In this section, we conducted a detailed error analysis to gain insights into the limitations and potential areas for improvement of the proposed model. Table 7 contains three error cases in the testing set. It should be noted that ScientificIE is a challenging task, and the proposed model exhibited common errors, as outlined in Example 1 (wrong entity type and relation type) or in Example 2 (incorrect spans). In certain instances, the proposed model successfully identified entities indicated in SemEval-2018 and not in SciERC. However, it tended to overpredict the relationship between these entity pairs (which is not an inaccurate relationship), as shown in Example 3. Besides, we observed that the correct entity predictions were missing in the gold labels ("learners" in Example 4 , or "post level" and "blog level" in Example 5). The relations between entity pairs were then accurately identified. The proposed model demonstrates the capacity to cover all entities and their relationships, including the most challenging and ambiguous cases. In contrast, the gold label annotations may not always capture these complex instances accurately.

[6] In this experiment, we considered the whole entity and relation types of both datasets.

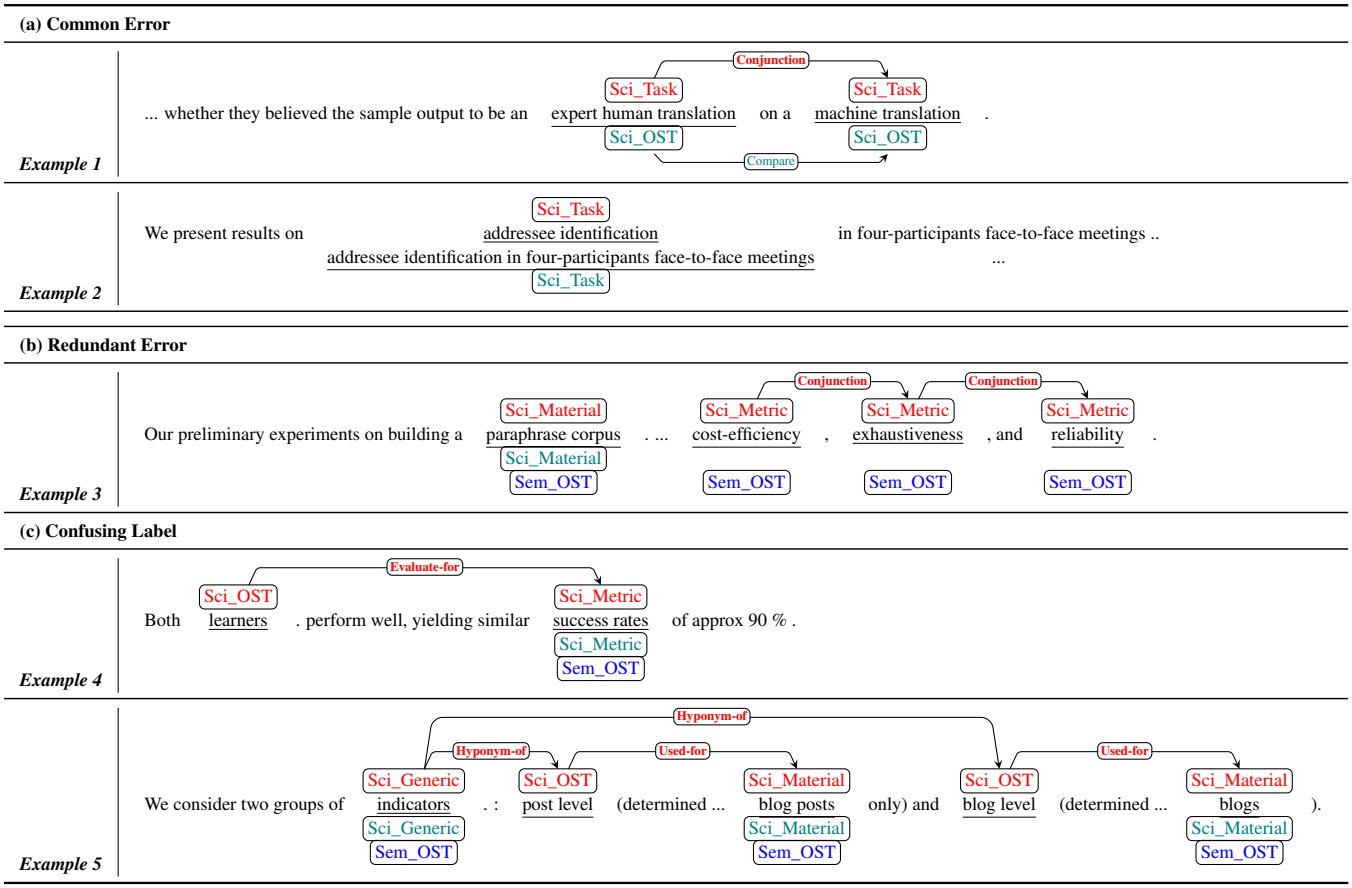


Table 7: The error samples are from the predictions from the proposed method: (a) the common sources of error, (b) predictions are redundant the relations, and (c) correct predictions are missing in the gold label. [*red] is predictions, [*green] is the gold label of SciERC, [*blue] is the gold label of SemEval-2018, OST is "OtherScientificTerm".

6 Related Work

ScientificIE systems can be developed using two main approaches: separate models, where entity extraction and relation extraction are treated as independent tasks with separate models trained for each (Xiao et al., 2020; Zhong and Chen, 2021; Ye et al., 2022), and joint models (end-to-end models) that tackle both tasks simultaneously (Eberts and Ulges, 2019; Luan et al., 2019; Santosh et al., 2021).

In recent years, various studies have prompted the community to explore innovative approaches to data labeling based on label variation (Passonneau et al., 2010; Plank et al., 2014; Basile et al., 2021; Gordon et al., 2021; Leonardelli et al., 2021; Prabhakaran et al., 2021; Uma et al., 2021a; Bassignana and Plank, 2022a; Plank, 2022). In this context, the proposed model leveraged variation labels in ScientificIE, thus demonstrating improved robustness with respect to label noise, in addition to higher performances in both tasks. To handle inconsistent labels and mitigate label noise, soft labeling techniques were introduced, such as the probabilistic soft labeling framework proposed by Fornaciari et al. (2021).

7 Conclusion

Label variation in ScientificIE introduces inconsistencies and ambiguities to labeled data, thus posing significant challenges for the training of accurate and reliable systems in this field. To overcome these challenges, we propose a multi-task learning approach that effectively handles label variations. By incorporating soft labels generated through multi-level agreements, we observed improvements in the performances demonstrated in entity and relation extraction tasks. The results indicate that label variations capture rich information and exhibit the potential to reduce data size requirements. Moreover, label variations are effective in handling ambiguous instances. The findings emphasize the significance of considering label variations in ScientificIE, and further promote its investigation in other domains and tasks.

Limitations

This study acknowledges several limitations that should be considered. First, the findings are based on a small dataset comprised of published research papers, which may limit the generalizability of the results to a larger population or different contexts. Second, the generation of accurate and reliable soft labels remains a challenge, as the manual setting of probability distributions introduces subjectivity. Additionally, the evaluation of the experiments solely relying on the F1-score using gold labels may be impacted by errors and inconsistencies within the gold label annotations, as revealed in the error analysis. Finally, it is essential to note that this work is primarily focused on the scientific domain, and the prevalence of conflict cases may differ in other domains, thus limiting the direct transferability of the findings.

Future research should address these limitations by incorporating larger and more diverse datasets, improving the methodology for generating soft labels, considering multiple evaluation metrics, and investigating the performances of large language models in ScientificIE tasks.

Acknowledgments

This work was (partly) supported by JSPS KAKENHI Grant Number 22K19818, and JST, AIP Trilateral AI Research, Grant Number JPMJCR20G9, Japan.

References

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pages 15–21. Association for Computational Linguistics.

Elisa Bassignana and Barbara Plank. 2022a. CrossRE: A cross-domain dataset for relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3592–3604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Elisa Bassignana and Barbara Plank. 2022b. What do you mean by relation extraction? a survey on datasets and study on scientific relation classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 67–83, Dublin, Ireland. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Markus Eberts and Adrian Ulges. 2019. Span-based joint entity and relation extraction with transformer pre-training. *The 24th European Conference on Artificial Intelligence (ECAI 2020)*.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana. Association for Computational Linguistics.

Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online. Association for Computational Linguistics.

Emily Jamison and Iryna Gurevych. 2015. Noise or additional information? leveraging crowdsource annotation item agreement for natural language tasks.

In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 291–297, Lisbon, Portugal. Association for Computational Linguistics.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3036–3046, Minneapolis, Minnesota. Association for Computational Linguistics.

Rebecca J. Passonneau, Ansaf Salleb-Aoussi, Vikas Bhardwaj, and Nancy Ide. 2010. Word sense annotation of polysemous words by multiple annotators. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

TYSS Santosh, Prantika Chakraborty, Sudakshina Dutta, Debarshi Kumar Sanyal, and Partha Pratim Das. 2021. Joint entity and relation extraction from scientific documents: role of linguistic information and entity types. *EEKE@ JCDL*, 21.

Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021a. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021b. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Chaojun Xiao, Yuan Yao, Ruobing Xie, Xu Han, Zhiyuan Liu, Maosong Sun, Fen Lin, and Leyu Lin. 2020. Denoising relation extraction from document-level distant supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3683–3688, Online. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4904–4917. Association for Computational Linguistics.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Learning with different amounts of annotation: From zero to many labels. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7620–7632, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

A Label Mapping

The datasets SemEval-2018 Task 7 and SciERC contain directly corresponding labels, as detailed in Section 2. To establish this correspondence, we compared the co-occurrence distribution of related relation labels between entity pairs in the 307 overlapping abstracts. To ensure consistency, we transferred entity labels with the same boundaries from SciERC, as SemEval-2018 Task 7 did not release entity types. With entity types only in SemEval, we retained type "OtherScientifiTerm_2". The co-occurrence score was computed using the following formula:

$$O(i, j, k) = \frac{A(e_i^1, e_j^2)^{r_k}}{N_i^1 + N_j^2} \tag{6}$$

where $A(e_i^1, e_j^2)^{r_k}$ denotes the number of occurrence relations r_k between entity pairs e_i^1 and e_j^2. N_i^1, N_j^2 represent the number of occurrences of entity label i, j was entity $1, 2$. The specific comparisons and descriptions can be found in Figures 4, 5, 6, 7, and 8. Corresponding relations exhibited similar co-occurrence distributions, with certain relations such as "Compare" and "Comparison" appearing most frequently between entity pairs such as "Method" and "Method" or "Task" and "Task".

B Common Relations

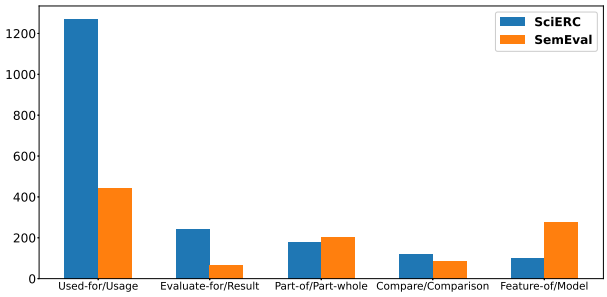


Figure 3: The distribution of common relations in the overlapped abstracts of the two datasets.

We observed that most relations in both datasets are labeled as *"Used-for/Usage"*. However, there is a notable disparity between the two datasets regarding label distribution. Specifically, in SemEval, labels such as *"Model"* and *"Part-whole"* have a significantly larger number of occurrences when compared with their counterparts in SciERC.

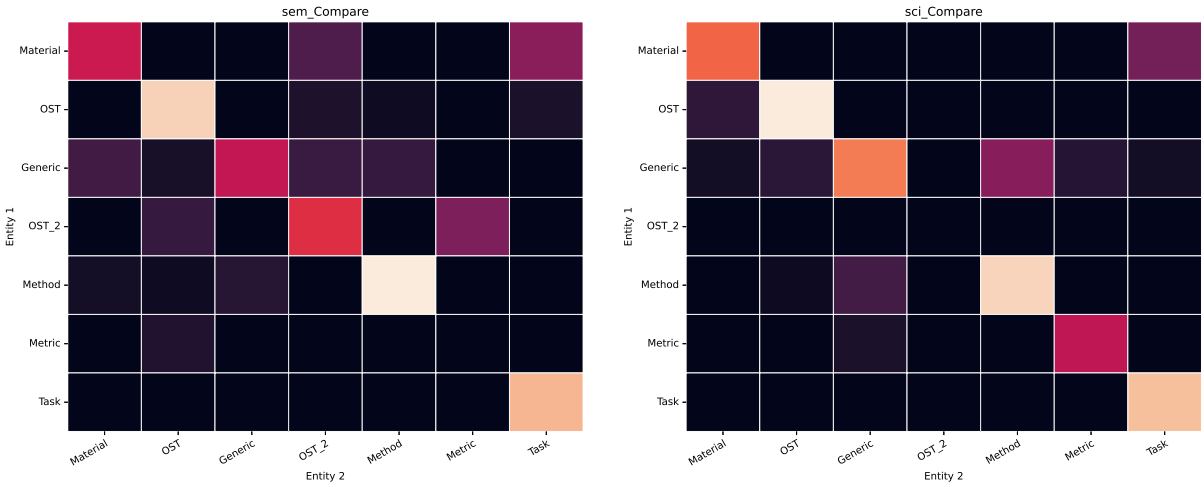


Figure 4: The co-occurrence distribution of “Compare/Comparision” in two overlapped corpora.

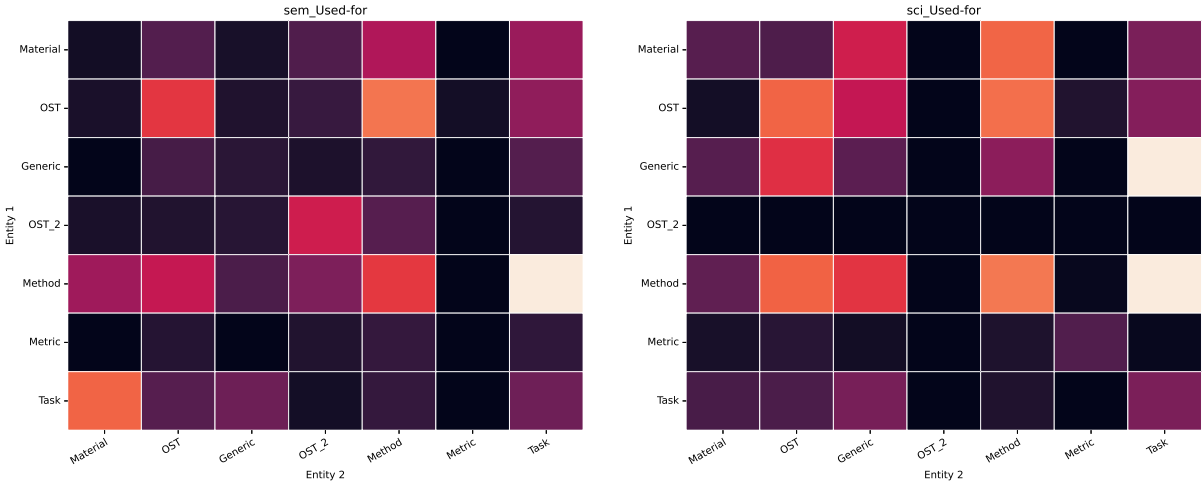


Figure 5: The co-occurrence distribution of “Used-for/Usage” in two overlapped corpora.

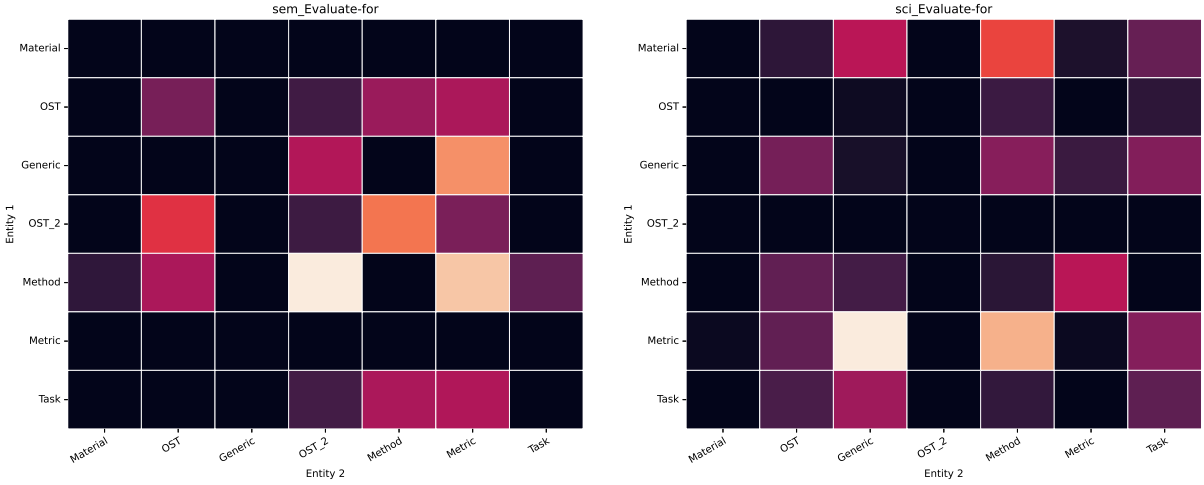


Figure 6: The co-occurrence distribution of “Evaluate-for/Result” in two overlapped corpora.

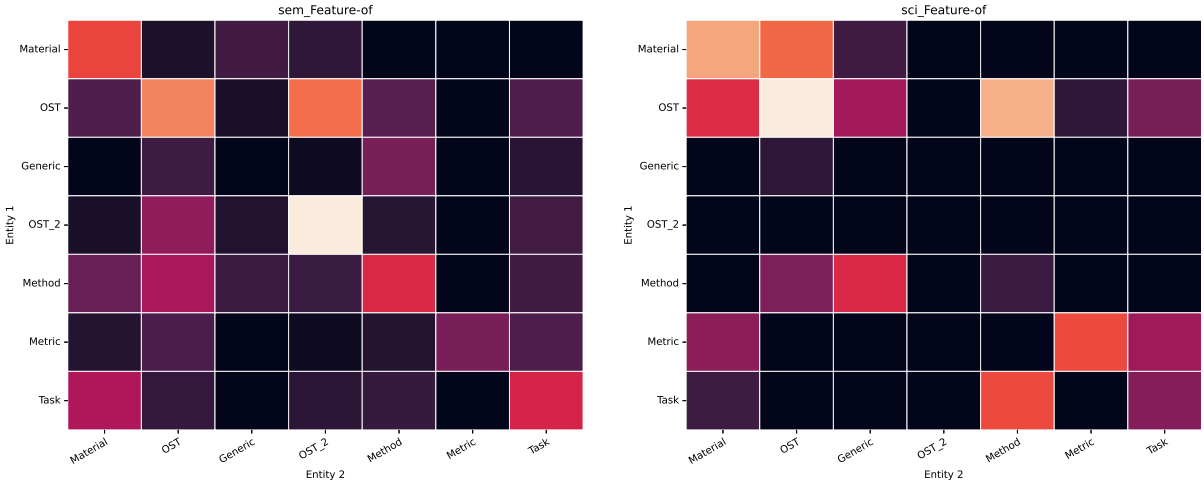


Figure 7: The co-occurrence distribution of "Feature-of/Model" in two overlapped corpora.

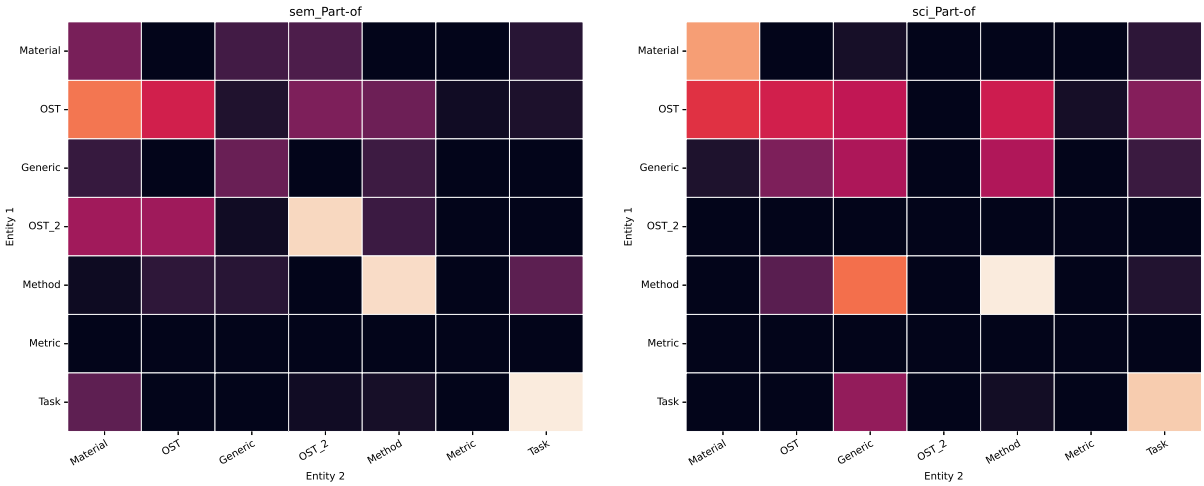


Figure 8: The co-occurrence distribution of "Part-of/Part-whole" in two overlapped corpora.