# 7

# S2ORC-SemiCause: Annotating and Analysing Causality in the Semiconductor Domain

**Xing Lan Liu[1], Eileen Salhofer[1,2], Anna Safont Andreu[3,4], and Roman Kern[2]**

[1]Know-Center GmbH, Austria
[2]Graz University of Technology, Austria
[3]University of Klagenfurt, Austria
[4]Infineon Technologies Austria

**Abstract**

For semiconductor manufacturing, easy access to causal knowledge documented in free texts facilitates timely Failure Modes and Effects Analysis (FMEA), which plays an important role to reduce failures and to decrease production cost. Causal relation extraction is the tasks of identifying causal knowledge in natural text and to provide a higher level of structure. However, the lack of publicly available benchmark causality datasets remains a bottleneck in the semiconductor domain. This work addresses this issue and presents the S2ORC-SemiCause benchmark dataset. It is based on the S2ORC corpus, which has been filtered for literature on semiconductor research, and consecutively annotated by humans for causal relations. The resulting dataset differs from existing causality datasets of other domain in the long spans of causes and effects, as well as causal cue phrases exclusive to the domain semiconductor research. As a consequence, this novel datasets poses challenges even for state-of-the-art token classification models such as S2ORC-SciBERT. Thus this dataset serves as benchmark for causal relation extraction for the semiconductor domain.

**Keywords:** causality, relation extraction, information extraction, bertology, annotation.

## 7.1 Introduction

Although causality represents a simple logical idea, it becomes a complex phenomenon when appearing in textual form. Natural language provides a wide variety of structures to represent causal relationships that can obfuscate the causal relations expressed via cause and effect. The task of causal relation extraction aims at extracting sentences containing causal language and identifying causal constituents and their relations [17].

In the last years significant progress have been made in automatizing the identification of causal cues and extraction of causal relation in natural language, defining it as a multi-way classification problem of semantic relationships [6], designing a lexicon of causal constructions [2, 3], and insights how to achieve high inter-rater agreement [13]. Approaches have been developed in scientific domains traditionally dominated by textual information, such as biomedical sciences. Here, models to process causal relations are facilitated and accelerated with the development of benchmark datasets such as BioCause [10]. Such datasets not only allow for comparison and automatic evaluation of custom causal extractors, but also allow for training high performing supervised models.

For semiconductor manufacturing, much of existing knowledge can be considered to be causal, highlighted by approaches like Ishikawa causal diagrams as well as the Failure Modes and Effects Analysis (FMEA) tool which captures root causes of potential failures. Even though such FMEA document provides more structure than natural language text, dedicated pre-processing is required before further processing [12]. A signification amount of such causal knowledge is captured in textual documents, such as reports and knowledge bases. However, there is no publicly available annotated dataset for causal relation extraction yet. As a consequence, in this work we propose such a dataset, named *S2ORC-SemiCause*. The source for the documents of this novel dataset is the S2ORC academic corpus, which has been filtered for documents of relevance for the semiconductor domain. Human annotators identified causal cues and causal relations in the documents of the corpus. To achieve consistent and reproducible results, an annotation guideline was created and the annotation processes was conducted in multiple phases. To provide baseline performance, the pre-trained language model BERT [1], which is currently considered state of the art for many natural language processing (NLP) tasks was adapted for the task. An error analysis gives insights on the challenges of future causal relation extraction methods.

In summary, our main contributions are:

- *S2ORC-SemiCause*, a causality dataset for the semiconductor domain that aims to provide a benchmark for causal relation extraction performances and facilitate research on dedicated methods;
- Practical annotation guidelines designed to yield high inter-annotator agreement for semiconductor literature, to enable the creation of further, similar datasets;
- Identified the key differences of *S2ORC-SemiCause* compared to other domains, and highlighted the resulting challenges for state-of-the-art NLP models.

## 7.2 Dataset Creation

### 7.2.1 Corpus

Our semiconductor corpus is selected from the 24 million papers in the engineering and related domains from the S2ORC corpus [8] (total 81.8 million papers). The subdomain is further filtered using a series of keywords specific for the semiconductor domain, such as device locations, electrical and physical faults, technologies (e.g. SFET), Focused Ion Beam, etc. For a paper to be selected, it needs to include at least four of these keywords.

From the resulting subset of 21 thousand papers, 400 abstract and 400 paragraphs are randomly sampled, among which 600 sentences are selected randomly for annotation.

### 7.2.2 Annotation Guideline

We have adapted the annotation guidelines[1] from the creation of BECauSE Corpus 2.0 [3]. The main differences are (1) the relation types "Motivation" and "Purpose" are further merged into one type (name "Purpose") since it is found from previous work [5] that annotators have difficulty distinguishing these two types; (2) *"max-span" rule*, namely, the span should include full phrase or clause. The *"max-span"* rule not only retains important context information for the causal relations, but also enables higher inter-annotator agreement. This was also motivated that it assumed to be easier to automatically reduce a phrase to its heads, instead of expanding a short, existing annotation.

[1] The annotation guideline will be make public at https://github.com/tugraz-isds/kd.

**Table 7.1** Inter-annotator agreement for the first two iterations. *Arg1* (cause) refers to the span of the arguments that lead to *Arg2* (effect) for the respective relation type.

| | **Iteration 1** | **Iteration 2** |
|---|---|---|
| Relation classification Cohen's $\kappa$ | 0.65 | 0.80 |
| Consequence Arg1 $F_1$ | 0.55 | 0.71 |
| Consequence Arg2 $F_1$ | 0.60 | 0.81 |
| Purpose Arg1 $F_1$ | 0.00 | 0.92 |
| Purpose Arg2 $F_1$ | 0.00 | 0.80 |
| $F_1$ micro average | 0.49 | 0.78 |

**Table 7.2** Comparison of labels generated by both annotators for Iteration 2. Examples and total counts (in number of arguments) for each type also given. Arg1 and Arg2 are highlighted with blue and yellow background, respectively. Partial overlapped texts are highlighted with green background.

| **Type** | **#** | **Example sentence** |
|---|---|---|
| Exact match | 54 | *In fact, and for the soil in question, the capillary rise process is low , so the indirectly loss by evaporative loss is low too .* |
| Partial overlap | 8 | *This result suggests a possible dynamical influence of the mesospheric layers on the lower atmospheric levels .* |
| Only one annotator | 14 | *The wing displaces away from the ground , as a result of the reduction in (-ve) lift .* |

### 7.2.3 Annotation Methodology

Since the annotations should contain as little ambiguity as possible, we aimed to design a methodology to achieve consistent annotations. To this end, the dataset was annotated in a total of 3 iterations. For the first two iterations with 50 sentences each, both annotators label the same set, so that inter-annotator-agreement (IAA) can be evaluated. Between the two iterations, the two annotators discussed the results and updated the guideline.

Table 7.1 shows that there are significant improvement in Inter-Annotator Agreement (IAA) from iteration 1 to iteration 2, both in terms of Cohen's $\kappa$, and $F_1$. The main improvement comes from (1) direction for *Purpose* relation (namely, *arg2* should be the purpose); (2) *"max-span" rule*, namely, the span should include full phrase or clause.

With Iteration 2, the two annotators reached a substantial agreement, where both Cohen's $\kappa$ for relation classification and $F_1$ for argument spans are around 0.8. For reference, in Dunietz et al. [3] a Cohen's $\kappa$ of 0.70 was reported for the relation type. Results of detailed inspection are summarized

**Table 7.3 Descriptive statistics of benchmark datasets**. Overview of CoNLL-2003 (training split) and BC5CDR (training split) for named entity recognition, as well as causality dataset BioCause (full dataset), and S2ORC-SemiCause (training split).

| | **CoNLL-2003** | **BC5CDR** | **BioCause** | **S2ORC-SemiCause** |
|---|---|---|---|---|
| #sentences | 14,042 | 4,612 | 37,422 | 360 |
| Avg. sentence length (in tokens) | 14.5 | 25.0 | 7.8 | 32.0 |
| Avg. argument length (in tokens) | 1.4 | 1.5 | 3.6 | 9.5 |

in Table 7.2. For 54 arguments, both annotators agree in both span and argument type. The remaining disagreements are from (1) one annotator misses a relation (14 occurrences); (2) only partial overlap of the annotated spans by both annotators (8 occurrences).

Based on the insights from the updated baseline, the first set of document was revisited and both set of annotations from the first two iterations were then merged manually. In addition, for the 3rd iteration, two extra sets of 250 sentences were annotated by each annotators. As a result, our dataset consist of 600 sentences annotated with Consequence and Purpose relations.

### 7.2.4 Dataset Statistics

We notice that compared to other benchmark NER datasets, such as CoNLL2003 [4], BC5CDR [7], and BioCause [10] (see Table 7.3), S2ORC-SemiCause dataset differs in terms of (1) smaller size; (2) longer sentence length; (3) longer argument length. While data size is found to be generally sufficient for entity recognition tasks [14], and longer sentence length is found to be preferred [14], the effect of longer argument length remains to be evaluated.

### 7.2.5 Causal Cue Phrases

When present, the causal cue phrases are also annotated. Figure 7.1 depicts the most common cue phrases for both relation types. *"To"* is the most frequently occurring cue because it is by far the most dominating cue phrase for relation type *purpose*. The cue phrases for *consequence* are much more diverse. Compared to other corpus of general domain [9, 11], in S2ORC-SemiCause dataset, cue words such as *increase, decrease, improve, reduce* are also ranked very high.

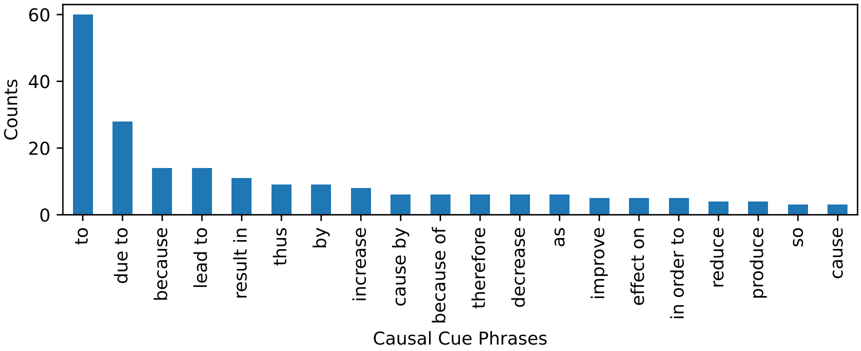

**Figure 7.1 Causal cue phrases ranked by frequency for all sentences in S2ORC-SemiCause dataset**.

## 7.3 Baseline Performance

To establish a point of reference for the community, we provide an initial baseline performance. For the baseline approach we considered the causal relation extraction task as an sequence classification task. As a technical realisation, we fine-tuned BERT on the down-stream task of token-level classification [1]. An error analysis is then performed to identify the main challenges in extracting causal relations from scientific publications in semiconductor research.

### 7.3.1 Train-Test Split

The total 600 sentences are split into training, validation, and test sets, with the ratio 60 : 20 : 20, stratified on relation type[2]. In addition, also the iterations were stratified evenly to avoid unwanted biases. The descriptive statistics for each split is listed in Table 7.4.

### 7.3.2 Causal Argument Extraction

As recommended in [1], which describes a similar scenario, we considered the task as a token-level classification. Namely, a pretrained BERT model is stacked with a linear layer on top of the hidden-states output, before fine-tuned on training examples. And the pretrained S2ORC-SciBERT model [8] is selected for fine-tuning using transformers library from Hugging Face [16].

[2] We release all data for future studies at https://github.com/tugraz-isds/kd

**Table 7.4** Descriptive statistics of *S2ORC-SemiCause* dataset. *#-sent*: total number of annotated sentences, *#-sent no relations*: number of sentences without causality, *Argument*: total amount and mean length (token span) of all annotated argument, *Consequence/Purpose*: amount and mean length of cause and effect arguments for the respective relation types.

| | **#-sent** | **#-sent no relations** | **Argument** | | **Consequence** | | | | **Purpose** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | **cause** | | **effect** | | **cause** | | **effect** | |
| | | | **count** | **mean** | **count** | **mean** | **count** | **mean** | **count** | **mean** | **count** | **mean** |
| overall | 600 | 291 | 670 | 9.4 | 258 | 8.4 | 290 | 9.2 | 58 | 10.8 | 64 | 12.9 |
| train | 360 | 174 | 405 | 9.5 | 155 | 8.5 | 178 | 9.1 | 34 | 11.1 | 38 | 13.5 |
| dev | 120 | 55 | 122 | 9.3 | 49 | 8.1 | 52 | 8.8 | 10 | 9.7 | 11 | 16.1 |
| test | 120 | 62 | 143 | 9.3 | 54 | 8.3 | 60 | 9.9 | 14 | 10.9 | 15 | 8.9 |

**Table 7.5** Baseline performance using BERT with a token classification head. Both the $F_1$ scores and the standard derivation over 7 different runs are shown. Despite the small sample size, the standard deviation remain low, similar to previous work [14].

| **Relation** | **Argument** | **#** | **$F_1$** | **$F_1$-filter** | **$F_1$-filter partial** |
|---|---|---|---|---|---|
| Consequence | Arg1 | 54 | 0.43 ± 0.03 | 0.48 ± 0.02 | 0.59 ± 0.01 |
| Consequence | Arg2 | 60 | 0.45 ± 0.03 | 0.50 ± 0.03 | 0.62 ± 0.02 |
| Purpose | Arg1 | 14 | 0.20 ± 0.07 | 0.25 ± 0.10 | 0.50 ± 0.05 |
| Purpose | Arg2 | 15 | 0.31 ± 0.06 | 0.36 ± 0.08 | 0.57 ± 0.07 |
| micro average | | 143 | 0.39 ± 0.02 | 0.45 ± 0.02 | 0.59 ± 0.01 |

The resulting $F_1$ scores[3] are shown in Table 7.5 and is remarkable lower than for other benchmark NER datasets when down-sampled to similar size [14].

### 7.3.3 Error Analysis

In order to understand the causes for the low $F_1$ score of the baseline model, an error analysis is performed.

#### Length of Argument Span

Firstly, a manual inspection revealed that for 30 ± 4 (out of the total 120) sentences, the fine-tuned model predicts sequences similar to [O I I ··· ], i.e., the models did not learn that an argument must always start with a "B" type with the IOB (Inside–Outside–Beginning) notation.

We hypothesize that this might be because our argument spans are much longer than other datasets (see Table 7.4 and Table 7.3). As a result, either the self-attention might no longer efficiently keep track of the [B I ···] pattern, or the over-abundant "I" class might bias the model loss.

[3] The best performance is found using learning rate $1.5e - 4$, batch size 8, warm up steps 10, and 10 epochs.

**Table 7.6** Comparison of predicted and annotated argument spans for the test split. Examples and total counts (in number of arguments) for correct prediction and for each error source are also given. Arg 1 and Arg 2 are highlighted with blue and yellow background, respectively. Partial overlapped texts are highlighted with green background.

| Type | # | Example sentence |
|---|---|---|
| Exact match | 68 | *The broad peak at 5 eV is due to N(2p) electrons.* |
| Partial overlap | 41 | *These safe zones are provided to a model predictive controller as reference to generate feasible trajectories for a vehicle.* |
| Spurious | 46 | *The roles of initial concentrations, space dimension and ratio of the reactant diKusinties in the modification of the reaction rate by many - particle eMects are compared with computer simulations.* |
| Missed | 34 | *This result validates the bolometric IR luminosities derived from MIR luminosities.* |

Following this hypothesis, we expect better performances for shorter arguments than for longer. Indeed we observe that correct predictions are shorter by 2.7 tokens on average ($p\_value = 0.008$).

To quantify the effect of such incorrect $[\mathrm{O}\ \ \mathrm{I}\ \ \mathrm{I}\ \cdots]$ sequences, we re-evaluated $\mathrm{F}_1$ score after filtering out such predictions. The results are shown in Table 7.5 as "$\mathrm{F}_1$-filter", and an improvement of 6 points is observed compared to the $\mathrm{F}_1$ score before filtering.

### Predictions with Partial Overlap

Out of the predicted argument, 41 were counted as incorrect, but overlapped partially (see example in Table 7.6), and manual inspection suggest that they often contain valid causal information.

Following [15], the model performance can be evaluated taking into account partial overlaps. The results are listed in Table 7.5 as "F1-filter partial", and the average F1 score becomes 0.59, which is about 80% of human performance (inter-annotator F1 value of 0.78), and is inline with the sample-size scaling as reported previously [14].

### Spurious and Missed Predictions

Spurious examples (false positives) are the cases where the model predicts a relation while annotators do not label. After manual inspection, we find it arguable that some *spurious* predictions made by the model might actually be valid causal relations as well. For example, the spurious example shown in Table 7.6 is arguably causal as well following the (*The role of ... in ...*) construct.

Missed examples (false negatives) are the cases where annotators have labelled while the model fails to predict a relation. For example, the missed example shown in Table 7.6 uses the rare causal trigger *derived from*, which might be the reason why the model failed to recognize.

## 7.4 Conclusions

Causality is critical knowledge in semiconductor manufacturing. In order to enable automatic causality recognition, we created the *S2ORC-SemiCause* dataset by annotating 600 sentences with 670 arguments for causal relation extraction from a subset of semiconductor literature taken from the S2ORC dataset. This unique dataset challenges established state-of-the-art techniques, because of its long spans for each argument. This benchmark dataset is intended to spur further research, fuel development of machine learning models, and to provide benefit to the NLP research in semiconductor domain.

## Acknowledgements

The research was conducted under the framework of the ECSEL AI4DI "Artificial Intelligence for Digitising Industry" project. The project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826060. The Know-Center is funded within the Austrian COMET Program–Competence Centers for Excellent Technologies under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG. We acknowledge useful comments and assistance from our colleagues at Know-Center and at Infineon.

## References

[1] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm):4171–4186, 2019.

[2] J. Dunietz, L. Levin, and J. Carbonell. Annotating causal language using corpus lexicography of constructions. *The 9th Linguistic Annotation Workshop held in conjunction with NAACL 2015*, (2014):188–196, 2015.

[3] J. Dunietz, L. Levin, and J. G Carbonell. The because corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104, 2017.

[4] E. F. Tjong Kim Sang, and F. De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003.

[5] D. Gaerber. Causal information extraction from historical german texts, 2022.

[6] I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proc. of the 5th Int. Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden, 2010. Association for Computational Linguistics.

[7] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu. Biocreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

[8] K. Lo, L. L. Wang, M. Neumann, R. Kinney, and D. Weld. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online, July 2020. Association for Computational Linguistics.

[9] Z. Luo, Y. Sha, K. Q. Zhu, S. W. Hwang, and Z. Wang. Commonsense causal reasoning between short texts. *Proc. Int. Workshop Tempor. Represent. Reason.*, pages 421–430, 2016.

[10] C. Mihăilă, T. Ohta, S. Pyysalo, and S. Ananiadou. BioCause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics*, 14, 2013.

[11] S. Pawar, R. More, G. K. Palshikar, P. Bhattacharyya, and V. Varma. Knowledge-based Extraction of Cause-Effect Relations from Biomedical Text. 2021.

[12] H. Razouk and R. Kern. Improving the consistency of the failure mode effect analysis (fmea) documents in semiconductor manufacturing. *Applied Sciences*, 12(4), 2022.

[13] I. Rehbein and J. Ruppenhofer. A new resource for German causal language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5968–5977, Marseille, France, May 2020. European Language Resources Association.

[14] E. Salhofer, X. L. Liu, and R. Kern. Impact of training instance selection on domain-specific entity extraction using bert. In *NAACL SRW*, 2022.

[15] I. Segura-Bedmar, P. Martínez, and M. Herrero-Zazo. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Proc. of the 7th Int. Workshop on Semantic Evaluation (SemEval 2013)*, 2013.

[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. v. Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-art natural language processing. In *Proc. of the 2020 Conf. on Empirical Methods in NLP: System Demonstrations*, 2020.

[17] J. Yang, S. C. Han, and J. Poon. A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, pages 1–26, 2022.