CiteWorth: Cite-Worthiness Detection for Improved Scientific Document Understanding

Anonymous NAACL-HLT 2021 submission

Abstract

Scientific document understanding is challenging as the data is highly domain specific and diverse. However, datasets for tasks with scientific text require expensive manual annotation and tend to be small and limited to only one or a few fields. At the same time, scientific documents contain much structure, which can potentially be used to build large labelled datasets. Given this, we explore the task of cite-worthiness detection in English, where a sentence is labelled for whether or not it cites an external source, by building CiteWorth, a rigorously cleaned labelled dataset from a large corpus of parsed scientific documents. We show that cite-worthiness detection is a challenging problem in and of itself; is a good test bed for domain adaptation; and that language model fine-tuning with cite-worthiness as a secondary task can lead to improved performance on downstream tasks in scientific document understanding. Moreover, we release CiteBert, a large pre-trained Transformer model fine-tuned on language modeling and citation detection on CiteWorth for improved scientific document understanding.

1 Introduction

Building effective natural language processing systems from scientific text is challenging due to the highly domain-specific and diverse nature of scientific language, and a lack of abundant sources of labelled data to capture this. While large scale repositories of parsed structured unlabelled text have recently been introduced (Lo et al., 2020), most datasets for downstream tasks such as named entity recognition (Li et al., 2016) and citation intent classification (Cohan et al., 2019) remain limited in size and highly domain specific. This begs the question: what useful training signals can be automatically extracted from massive unlabelled scientific text corpora to help improve scientific document understanding?

Scientific documents contain much inherent structure (sections, tables, equations, citations, etc.), which can facilitate creating large labelled datasets. Some recent examples include using paper field (Beltagy et al., 2019), the section to which a sentence belongs (Cohan et al., 2019), and the cite-worthiness of a sentence (Cohan et al., 2019; Sugiyama et al., 2010) as a training signal.

Cite-worthiness detection is the task of identifying *citing sentences*, i.e. sentences which contain a reference to an external source. It has useful applications, such as in assistive document editing, and as a first step in citation recommendation (Färber et al., 2018b). In addition, cite-worthiness has been shown to be useful in helping to improve the ability of models to learn other tasks (Cohan et al., 2019), making it an attractive topic to study in the context of building a large labelled dataset from scientific text. We also hypothesize that there is a strong domain shift between how different fields use citations, and that such a dataset is useful for studying domain adaptation problems with scientific text.

However, constructing such a dataset to be of high quality is surprisingly non-trivial. Building a dataset for cite-worthiness detection involves extracting sentences from a scientific document, labelling whether each sentence contains a citation, and removing all citation markers. As a form of distant supervision, this naturally comes with the hazard of adding spurious correlations, such as misparsed sentences or hanging punctuation, which can trivially indicate a cite-worthy or non-cite-worthy sentence. One can mitigate such noise in distant supervision either by designing one's model with this in mind (Riedel et al., 2010) or by taking extra care when curating the data (Takamatsu et al., 2012). In this work we opt for the latter, presenting CiteWorth, an iteratively and rigorously curated dataset for cite-worthiness detection in English, offering the dataset to the research community to facilitate further research on cite-worthiness detection.

Using CITEWORTH, we ask the following primary research questions:

> **RQ1**: How can a dataset for cite-worthiness detection be automatically curated with low noise (§3)?
>
> **RQ2**: Is such a dataset a good test bed for studying domain adaptation (§5)?
>
> **RQ3**: Can large scale cite-worthiness data be used to perform transfer learning to downstream scientific text tasks (§6)?

We demonstrate that CITEWORTH is of high quality through a manual evaluation, finding that it can be very useful for studying domain adaptation with large differences in how models generalize to data from different fields. Additionally, we find that cite-worthiness is a useful task for transferring to downstream scientific text tasks, in particular citation intent classification, for which we offer performance improvements over the current state-of-the-art model SciBERT (Beltagy et al., 2019).

In sum, our **contributions** are as follows:

- CITEWORTH, a dataset of 1.2M rigorously cleaned sentences from scientific papers labelled for cite-worthiness, balanced across 10 diverse scientific fields.
- A thorough analysis of the problem of cite-worthiness detection, showing that it is non-trivial as a problem in itself and can be a useful test bed for domain adaptation.
- New state of the art on citation intent detection via transfer learning from joint citation detection and language model fine-tuning on our data, with improved performance over SciBERT on several other tasks.

2 Related Work

2.1 Cite-Worthiness Detection

Cite-worthiness detection is the task of identifying *citing sentences* i.e. sentences which contain a reference to an external source. The reasons for citing are varied, e.g. to give credit to existing ideas or to provide evidence for a claim being made. Sugiyama et al. (2010) perform cite-worthiness detection using SVMs with features such as unigrams, bigrams, presence of proper nouns, and the classification of previous and next sentences. They create a dataset from the ACL Anthology Reference corpus (ACL-ARC, Bird et al. (2008)), using heuristics to remove citation markers. Färber et al. (2018b) document the performance of convolutional recurrent neural nets on a larger set of three datasets coming from ACL-ARC, arXiv CS (Färber et al., 2018a), and Scholarly Dataset 2[1]. Datasets from these studies suffer from high class imbalance, are limited to only one or a few domains, and little analysis into the datasets is given in order to understand the quality of the data or what aspects of the problem are difficult or easy. Additionally, these data are limited to only one or a few fields.

In addition to being a useful task in itself, cite-worthiness detection is useful for other tasks in scientific document understanding. In particular, it has been shown to help improve performance on the closely related task of citation intent classification (Jürgens et al., 2018) when used as an auxiliary task in a multi-task setup (Cohan et al., 2019). However, cite-worthiness detection has not been studied in a transfer learning setup as a pre-training task for multiple scientific text problems. In this work, we seek to understand to what extent cite-worthiness detection is a transferable task.

Scientific Document Understanding Numerous problems related to scientific document understanding have been studied previously. Popular tasks include named entity recognition (Li et al., 2016; Kim et al., 2004; Doğan et al., 2014; Luan et al., 2018), relation extraction (Kringelum et al., 2016; Luan et al., 2018), keyphrase extraction (Augenstein et al., 2017), dependency parsing (Kim et al., 2003), citation intent classification (Jürgens et al., 2018; Cohan et al., 2019), and fact checking (Wadden et al., 2020).

Datasets for scientific document understanding tasks tend to be limited in size and restricted to only one or a few fields, making it difficult to build models with which one can study cross-domain performance and domain adaptation. Here, we curate a large dataset of cite-worthy sentences spanning 10 different fields, showing that such data is both useful for studying domain adaptation and for transferring to related downstream scientific document understanding tasks.

3 RQ1: CITEWORTH Dataset Construction

The first research question we ask is: How can a dataset for cite-worthiness detection be automati-

[1] http://www.comp.nus.edu.sg/~sugiyama/SchPaperRecData.html

Biology
Wood Frogs (Rana sylvatica) are a charismatic species of frog common in much of North America. They breed in explosive choruses over a few nights in late winter to early spring. *The incidence in Wood Frogs was associated with a die-off of frogs during the breeding chorus in the Sylamore District of the Ozark National Forest in Arkansas* ~~(Trauth et al., 2000)~~*.*

Computer Science
Land use or cover change is a direct reflection of human activity, such as land use, urban expansion, and architectural planning, on the earth's surface caused by urbanization ~~[1].~~ Remote sensing images are important data sources that can efficiently detect land changes. *Meanwhile, remote sensing image-based change detection is the change identification of surficial objects or geographic phenomena through the remote observation of two or more different phases* ~~[2]~~*.*

Table 1: Excerpts from training samples in CITEWORTH from the Biology and Computer Science fields. Green sentences are cite-worthy sentences, from which citation markers are removed during dataset construction.

cally curated with low noise? To answer this, we start with the S2ORC dataset of parsed scientific articles (Lo et al., 2020). The dataset consists of data from 81.1M English scientific articles, with full structured text for 8.1M articles. The data also includes rich metadata, e.g. Microsoft Academic Graph (MAG) categories, linked citations, and linked figures and tables. It constitutes a massive repository of machine readable full text articles from which a large and high-quality cite-worthiness dataset may be built.

3.1 Data Filtering

Given the size of S2ORC, we start by reducing the candidate set of data to only articles which we are confident have been parsed with high quality. The criteria that we use to filter the papers are:

- The paper has a parsed abstract.
- The paper has a fully parsed PDF.
- The body text of the paper has been parsed.
- The bibliography of the paper has been parsed.
- The tables and figures of the paper have been parsed.
- The paper has venue information available.
- The paper has inbound citations.
- The paper belongs to at least one Microsoft Academic Graph category.

Filtering based on these criteria results in 5,494,387 candidate papers from which to construct our dataset. After filtering the candidate set of papers, we process the data as follows.

1. Select only sentences with parenthetical author-year or bracketed-numerical citation spans.
2. Select only sentences with citation spans at the end of a sentence.
3. Check if each possible citation span has been parsed by S2ORC.
4. Check if citation markers are left behind after removing citation spans e.g. hanging prepositions and punctuation.
5. Check if sentences begin with a capital letter and end with a period, question mark, or exclamation point to mitigate sentence tokenizer errors.
6. Check if each sentence is at least 20 characters long to mitigate sentence tokenizer errors.
7. If any sentence within a paragraph does not meet the above criteria, discard all sentences in the paragraph.

With the first two criteria, we limit the scope of cite-worthy sentences to being only those whose citation span comes at the end of a sentence, and whose citation format is parenthetical author-year form or bracketed-numerical form. In other words, cite-worthy sentences in our data are limited to those of the following forms.

> This result has been shown in previous work (Author et al., ####, Author2 et al. ...).
>
> This result has been shown in previous work [#-#].

In this, we ignore citation sentences which contain inline citations, such as "The work of Authors et al. (####) has shown this in previous work" and its counterpart with bracketed citations, as well as any sentence with a citation format that does not match the two we have selected.

Limiting cite-worthy sentences as such helps prevent spurious correlations in the data. Removing citations in the middle of a sentence runs the risk of rendering the sentence ungrammatical, providing a signal to machine learning models. While there are cases where inline citations could potentially be removed in their entirety and not destroy the sentence structure, this is beyond the scope of this paper and left to future work.

3.2 Extracting Cite-Worthy Sentences

Our dataset construction pipeline for a given paper begins by first extracting all paragraphs from the parsed body text which come from a limited list of permissible section titles. The full list is provided in Appendix A. The titles include common section titles such as "Introduction," "Methods," and "Discussion." Additionally, we check the citation format of the parsed citation spans to ensure they are of the the two forms we allow via regular expressions (see Appendix B).

After extracting all of the paragraphs from each permissible section, we iterate through each of them and extract cite-worthy and non-cite-worthy sentences. For a given paragraph, we first word and sentence tokenize the text using SciSpacy (Neumann et al., 2019). Each sentence is then checked for containing citations using the provided citation spans in the S2ORC dataset. In some cases, the sentence contains citations which were missed by S2ORC; these are checked using regular expressions, and if a match is found then the entire paragraph is ignored. Otherwise, the location and format of the citation is checked, again using regular expressions. If the citation is not at the end of the sentence, then the entire paragraph is ignored. Otherwise, we proceed to remove the citation text using the provided citation spans.

Simply removing the citation span runs the risk of adding spurious correlations in the data, such as with hanging punctuation and prepositional phrases e.g. "This was shown by the work of ~~Author et al. (####)~~." To mitigate this, we remove all hanging punctuation at the end of a sentence that is not a period, exclamation point, or question mark, and check for possible hanging citations using the regular expression provided in Appendix B. This regular expression is built iteratively as follows:

1. Run the pipeline and sample a dataset.
2. Train a SciBERT (Beltagy et al., 2019) model on this data and rank the most confident positives from a held out development set of data.
3. Observe the top 40 confident positives, incorporate any parser issues and hanging prepositions into the pipeline.

Metric	#
Total sentences	1,181,793
Total number of tokens	34,170,708
Train sentences	945,426
Dev sentences	118,182
Test sentences	118,185
Total cite-worthy	375,388
Total non-cite-worthy	806,405
Min char length	21
Max char length	1,447
Average char length	152
Median char length	142

Table 2: Various statistics of the CITEWORTH dataset.

In order to prevent bias in the negative data, we also check if any non-cite-worthy sentences match our regular expression. If any sentence fails this check, the entire sentence is ignored.

To mitigate issues with sentence parsing, we also ensure that the first character of each sentence is a capital letter, and that the sentence ends with a period, exclamation point, or question mark. If all criteria are met for all sentences in a paragraph, these sentences are added to the dataset. Finally, we build a dataset which is diverse across domains by evenly sampling paragraphs from the following 10 MAG categories, ensuring that each paragraph belongs to exactly one category: Biology, Medicine, Engineering, Chemistry, Psychology, Computer Science, Materials Science, Economics, Mathematics, and Physics. Example excerpts from the dataset are presented in Table 1, and the statistics for the final dataset are given in Table 2.

3.3 Manual Evaluation

In order to provide some measure of the general quality of CITEWORTH, we perform a manual evaluation of a sample of the data. We annotate the data for whether or not citation markers are completely removed, and for whether or not the sentences are well-formed, containing no obvious parsing or cleaning artifacts. We sample 500 cite-worthy sentences and 500 non-cite-worthy sentences randomly from the data. Additionally, we compare to a baseline where the only heuristic used is to remove citation spans based on the provided spans in the S2ORC dataset, which uses SCIENCEPARSE[2] to parse PDF documents and GROBID[3] to extract

[2] https://github.com/allenai/scienceparse

[3] https://github.com/kermitt2/grobid

Method	Parsed Correct	Markers Correct
Baseline	92.07	92.78
Ours	**98.90**	**98.10**

Table 3: Results of manually annotating 1000 random sentences (per method) from CITEWORTH and a naive baseline which only removes citations based on provided citation spans. "Parsed Correct" are results for correctly parsing the sentences, and "Markers Correct" are results for successfully removing citation markers. The data curated using our method has 6% fewer errors in terms of parsing and removal of citation markers, and less than 2% of the samples have some form of citation marker.

structured data from the text including citation spans. We again sample 500 cite-worthy and 500 non-cite-worthy sentences for annotation. The two sets are shuffled together and given to an independent annotator for labelling. The results for the manual annoatation can be seen in Table 3.

We see that the CITEWORTH data are of a much higher quality than removing citation markers based only on the citation spans. Overall, our heuristics improve on parse quality by 6.83% absolute and on removing markers of citations by 5.32% absolute. This results in 1.1% of the sample data containing parse issues, and 1.9% having trivial markers indicating a citation is present. We argue that this is a strong indicator of the quality of the data for the purpose of supervised learning.

4 Benchmarks

To characterize the difficulty of the problem and ability of models to learn the task of cite-worthiness detection, we run a variety of baseline models on CITEWORTH.

Logistic Regression As a simple baseline, we use a logistic regression model with TF-IDF input features. We use balanced class weights to mitigate the class imbalance, and perform randomized grid search to tune the hyperparameters of the model

Basic Transformer We additionally train a basic Transformer model from scratch (Vaswani et al., 2017), tuning the model hyperparameters on a subset of the training data via randomized grid search.

BERT We use a pretrained BERT model (Devlin et al., 2019) due to the strong performance of large pretrained Transformer models on downstream tasks.

Method	P	R	F1
Logistic Regression	46.65	64.88	54.28
Transformer	47.92	71.59	57.39
Bert Base	55.04	69.02	61.23
Scibert Base (no weighting)	**65.94**	51.62	57.91
Scibert Base	57.03	68.08	**62.06**
Scibert + PU	49.46	**82.12**	61.73

Table 4: F1 performance of baselines on the test set of CITEWORTH. Results are averaged across 5 seeds.

SciBERT SciBERT (Beltagy et al., 2019) is a BERT model pretrained on a large corpus of scientific text from Semantic Scholar (Ammar et al., 2018), and is therefore potentially better suited to fine-tuning on scientific cite-worthiness detection.

SciBERT + PU Learning Finally, we experiment with SciBERT trained using positive-unlabelled (PU) learning (Elkan and Noto, 2008) which has been shown to significantly improve performance on citation needed detection in Wikipedia and rumour detection on Twitter (Wright and Augenstein, 2020a).

Due to the imbalance in the distribution of classes, the loss for each of the models is weighted. For comparison, we include results for SciBERT without weighting the loss function. The results for our baseline models on the test set of the dataset are given in Table 4.

The best overall model is SciBERT using balanced class weights at an F1 score of 62.1. Using class weighting is highly important, resulting in an increase of almost 5 F1 points. Compared to not using class weights, PU learning performs significantly better, and leads to the highest recall of all models under test. Additionally, language model pre-training is useful, as both BERT and SciBERT perform significantly better than a transformer trained from scratch. Finally, pre-training on scientific text performs almost one F1 point better than pre-training on general text, indicating that domain specific pre-training is important.

To gain some insight into what the model learns, we visualize the most salient features from SciBERT for selected easy and hard examples in Figure 1. We use the InputXGradient method (Kindermans et al., 2016), specifically the variant using L2 normalization over neurons to get a pre-embedding score, as it has been recently shown to have the best overall agreement with human rationales versus several other explainability techniques (Atanasova et al., 2020). The model is able

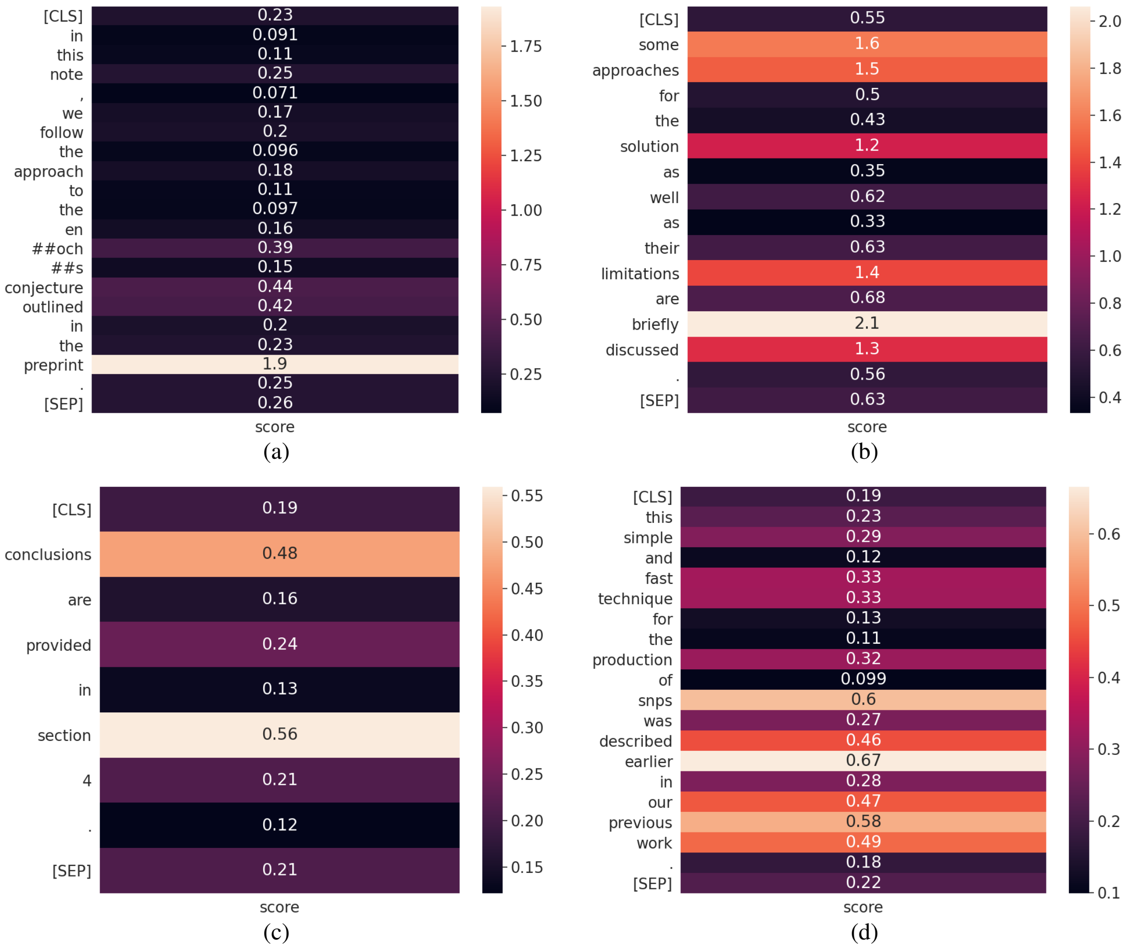

Figure 1: Visualization of explanations for incorrect and correctly classified examples from the dev set using InputXGradient (Kindermans et al., 2016). (a) Positive class explanations for a selected correctly classified cite-worthy sentence. (b) Negative class explanations for a cite-worthy example incorrectly classified as non-cite-worthy. (c) Negative class explanations for a correctly classified non-cite-worthy example. (d) Positive class explanations for a non-cite-worthy example incorrectly classified as cite-worthy. Positive class = cite-worthy, negative class = non-cite-worthy.

to pick up on obvious markers of cite-worthy and non-cite-worthy sentences, such as that a sentence refers to a preprint (Figure 1a) or to different sections within the paper itself (Figure 1c). We also see that the dataset contains many relatively difficult instances, for example with the model observing "briefly discussed" as an indicator that an instance is non-cite-worthy when it is in fact cite-worthy (Figure 1b). From this, we hypothesize that performance on this data can be improved by considering the context of a sentence, such as surrounding sentences or the whole document.

5 RQ2: Domain Adaptation

We next ask: is CiteWorth a good test bed for studying domain adaptation in scientific text? To answer this, we study the relationships between cite-worthiness data from different fields and how SciBERT performs in a cross-domain setup. For ease of analysis we limit the scope of fields to 5 of the 10 fields in the dataset which cover a wide spectrum: Chemistry, Engineering, Computer Science, Psychology, and Biology.

As a first step, we visualize the embedding space for data from each of these domains using the method of Aharoni and Goldberg (2020). In this, the data is passed through BERT and the output representations for each token in a sentence are average pooled. These representations are visualized in 2D space via PCA in Figure 2. It is clear that similar fields occupy closer space, with engineering and computer science sharing closer representations, as well as biology and chemistry. We perform clustering on this data using a Gaussian mixture model similarly to Aharoni and Goldberg (2020), finding that domains form somewhat distinct clusters with a cluster purity of 57.61. This demonstrates that the data in different fields are drawn from different distributions, making them amenable for use in studying domain adaptation.

To further characterize the relationship between field and domain, we perform a cross validation experiment using the 5 selected fields, training on one field and testing on another for all 25 combinations. The results for the 5x5 train/test setup using SciBERT are given in Table 5[4].

Not surprisingly, the best performance for each split occurs when training on data from the same field, with the exception of Engineering which performs approximately as well as Computer Science. We also observe high variance in the maximum performance for each field (σ = 4.46), as well as between different fields on the same test data despite large pretrained Transformer models such as SciBERT being relatively invariant across domains (Wright and Augenstein, 2020b). This suggests stark differences in the input distribution from each field, making field a good proxy for domain when studying domain adaptation. Additionally, we observe a slight inverse correlation between distance in the embedding space and performance on different domains (Pearson's correlation: -0.183, Spearman's ρ: -0.222). We hypothesize that this correlation would be more pronounced when not using a large pre-trained Transformer model, as was demonstrated by Aharoni and Goldberg (2020). Ultimately, we conclude that CITEWORTH is a good test bed for domain adaptation.

6 RQ3: Cite-Worthiness for Transfer Learning

The final question we ask is: to what extent is cite-worthiness detection transferable to downstream tasks in scientific document understanding? To answer this, we fine tune SciBERT on the task of cite-worthiness detection as well as masked language modeling (MLM) on CITEWORTH, followed by fine-tuning on several document understanding tasks. The tasks we evaluate on come from Beltagy et al. (2019) and are categorized as follows.

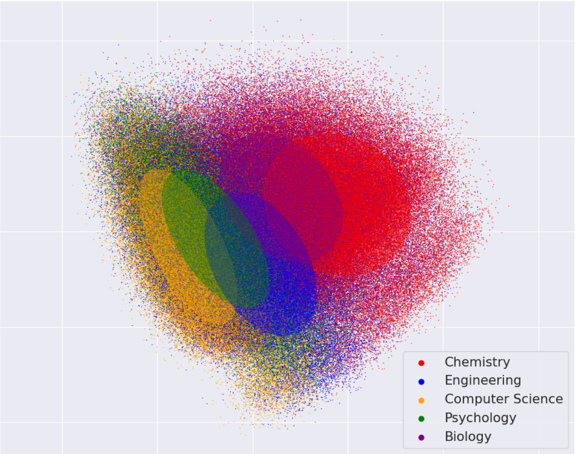

Figure 2: Visualizing the BERT embeddings for 5 of the 10 domains from CITEWORTH using the method by Aharoni and Goldberg (2020). Clustering is performed using Gaussian Mixture Models.

Train \ Test	Ch	E	CS	P	B
Ch	**63.52**	52.15	51.56	55.08	64.33
E	63.04	55.04	55.14	58.39	64.13
CS	62.29	**55.06**	**56.16**	58.56	63.81
P	62.12	53.21	52.53	**60.32**	65.24
B	63.06	52.62	52.41	58.71	**65.56**
σ	0.52	1.22	1.77	1.71	0.67

Table 5: F1 performance on different domain adaptation settings for the fields (Ch)emistry, (E)ngineering, (C)omputer (S)cience, (P)sychology, and (B)iology. Out-of-domain tests use the entire set of data from that field, while in domain tests use 80% of data for training, 10% for validation, and 10% for test.

- Named Entity Recognition (NER): This task involves labelling the spans of different types of entities in a document.
- PICO: This task is similar to NER, but for demarcating populations, interventions, what the interventions are compared to, and the outcomes of clinical trials.
- Relation Extraction (REL): This task involves labelling a sequence for the relationship between two eintities.
- Text classification (CLS): Finally, we test on several text classification tasks where the goal is to classify a sentence into one or more categories. The particular classification tasks we test are citation intent classification and paper field classification.

We compare four variants of pre-training and fine-tuning, given as follows.

[4] A table with results for the 10x10 setup with all fields is given in Appendix C.

Dataset	Reference	Task	Base	LM	Cite	LM + Cite
BC5CDR	Li et al. (2016)	NER	89.84 ± 0.18	**90.03 ± 0.11**	89.73 ± 0.25	90.02 ± 0.79
JNLPBA	Kim et al. (2004)	NER	77.02 ± 0.36	77.13 ± 0.53	76.97 ± 0.44	**77.15 ± 0.58**
NCBI-Disease	Doğan et al. (2014)	NER	**88.79 ± 0.35**	88.53 ± 0.58	88.66 ± 0.57	88.31 ± 0.43
SciERC	Luan et al. (2018)	NER	67.08 ± 0.50	66.64 ± 0.47	67.12 ± 0.46	**67.48 ± 0.45**
EBM-NLP	Nye et al. (2018)	PICO	76.61 ± 0.21	**76.69 ± 0.28**	76.55 ± 0.88	76.41 ± 0.32
ChemProt	Kringelum et al. (2016)	REL	83.17 ± 0.43	**83.26 ± 0.90**	82.70 ± 1.06	83.16 ± 0.63
SciERC	Luan et al. (2018)	REL	80.21 ± 0.81	**80.68 ± 1.04**	80.00 ± 1.73	80.58 ± 0.96
ACL-ARC	Jürgens et al. (2018)	CLS	71.82 ± 2.93	70.95 ± 2.25	**73.68 ± 2.75**	72.92 ± 3.76
SciCite	Cohan et al. (2019)	CLS	84.83 ± 0.65	85.18 ± 0.47	85.32 ± 0.16	**85.35 ± 0.29**
PaperField	Beltagy et al. (2019)	CLS	65.48 ± 0.18	**65.57 ± 0.27**	65.46 ± 0.24	65.42 ± 0.48
Average			78.386	78.466	78.619	**78.680**

Table 6: Performance on various downstream scientific document understanding tasks as presented by Beltagy et al. (2019). The metrics used are as follows: NER is span-level F1, PICO is token level F1, relation extraction is macro-F1, and ChemProt is micro-F1. All runs are averaged across 5 seeds.

Base SciBERT without fine tuning.

LM SciBERT with MLM fine tuning on CiteWorth.

Cite SciBERT fine-tuned for the task of cite-worthiness detection. The classifier is a pooling layer on top of the `[CLS]` representation of SciBERT, followed by a classification layer.

LM + Cite SciBERT with MLM fine tuning and cite-worthiness detection. The two tasks are trained jointly i.e. on each batch of training, the model incurs a loss for both MLM and cite-worthiness detection which are summed together.

The results for all experiments are given in Table 6. Note that the reported results for SciBERT are on re-running the model locally for fair comparison. We first observe that incorporating our dataset into fine-tuning tends to improve model performance across all tasks to varying degrees, with the exception of NER on the NCBI-Disease corpus. The tasks where cite-worthiness as an objective has the most influence are the two citation intent classification tasks (ACL-ARC and SciCite). We see average improvements of 1.8 F1 points for the ACL-ARC dataset (including 2 points F1 improvement over the minumum and maximum model performance SciBERT) and 0.5 F1 points on SciCite. The best average performance is from the model which incorporates both MLM and cite-worthiness as an objective, which we call CiteBert[5].

For other tasks, fine-tuning the language model on CiteWorth data tends to be sufficient for improving performance, though the margin of improvement tends to be minimal. CiteWorth is relatively small compared to the corpus on which SciBERT is originally trained (30.7M tokens for the train and dev splits on which we train versus 3.1B), so one could potentially see further improvements by incorporating more data or including cite-worthiness as an auxiliary task during language model pre-training. However, this is outside the scope of this work.

7 Conclusion

In this work, we present an in-depth study into the problem of cite-worthiness detection in English. We rigorously curate CiteWorth, a high-quality dataset for cite-worthiness detection; show that CiteWorth is a good testbed for studying domain adaptation in scientific text; and show that in a transfer-learning setup one can achieve state of the art results on the task of citation intent classification using this data. In addition to studying cite-worthiness and transfer learning, CiteWorth is suitable for use in downstream natural language understanding tasks. As we retain the S2ORC metadata with the data, one could potentially use the data to study joint cite-worthiness detection and citation recommendation. Additionally, as we extract all sentences at the paragraph level, the data could be used to study cite-worthiness as a structured prediction problem, which we plan to do in future work. We hope that the data and accompanying fine-tuned model will be useful to the research community working on problems in the space of scientific language processing.

[5] We plan to release the fine-tuned language model with cite-worthiness detection as a downloadable model in the HuggingFace model hub.

References

Roee Aharoni and Y. Goldberg. 2020. Unsupervised Domain Clusters in Pretrained Language Models. In *ACL*.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, et al. 2018. Construction of the Literature Graph in Semantic Scholar. *NAACL HLT 2018*, pages 84–91.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A Diagnostic Study of Explainability Techniques for Text Classification. In *Proceedings of EMNLP*. Association for Computational Linguistics.

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 Task 10: ScienceIE-Extracting Keyphrases and Relations from Scientific Publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555.

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 3606–3611.

Steven Bird, Robert Dale, Bonnie J Dorr, Bryan Gibson, Mark Thomas Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R Radev, and Yee Fan Tan. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural Scaffolds for Citation Intent Classification in Scientific Publications. 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 3586–3596.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI Disease Corpus: a Resource For Disease Name Recognition and Concept Normalization. *Journal of biomedical informatics*, 47:1–10.

Charles Elkan and Keith Noto. 2008. Learning Classifiers From Only Positive and Unlabeled Data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 213–220.

Michael Färber, Alexander Thiemann, and Adam Jatowt. 2018a. A High-Quality Gold Standard for Citation-Based Tasks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Michael Färber, Alexander Thiemann, and Adam Jatowt. 2018b. To Cite, or Not to Cite? Detecting Citation Contexts in Text. In *European Conference on Information Retrieval*, pages 598–603. Springer.

David Jürgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the Evolution of a Scientific Field through Citation Frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. GENIA Corpus—a Semantically Annotated Corpus for Bio-Textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the Bio-Entity Recognition Task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 70–75. Citeseer.

Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. 2016. Investigating the Influence of Noise and Distractors on the Interpretation of Neural Networks. *arXiv preprint arXiv:1611.07270*.

Jens Kringelum, Sonny Kim Kjaerulff, Søren Brunak, Ole Lund, Tudor I Oprea, and Olivier Taboureau. 2016. ChemProt-3.0: a Global Chemical Biology Diseases Mapping. *Database*, 2016.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983.

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.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J Marshall, Ani Nenkova, and Byron C Wallace. 2018. A Corpus With Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 197. NIH Public Access.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling Relations and Their Mentions Without Labeled Text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

Kazunari Sugiyama, Tarun Kumar, Min-Yen Kan, and Ramesh C Tripathi. 2010. Identifying Citing Sentences in Research Papers Using Supervised Learning. In *2010 International Conference on Information Retrieval & Knowledge Management (CAMP)*, pages 67–72. IEEE.

Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. 2012. Reducing Wrong Labels in Distant Supervision for Relation Extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 721–729.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

David Wadden, Kyle Lo, Lucy Lu Wang, Shanchuan Lin, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *EMNLP*.

Dustin Wright and Isabelle Augenstein. 2020a. Claim Check-Worthiness Detection as Positive Unlabelled Learning. In *Findings of EMNLP*. Association for Computational Linguistics.

Dustin Wright and Isabelle Augenstein. 2020b. Transformer Based Multi-Source Domain Adaptation. In *Proceedings of EMNLP*. Association for Computational Linguistics.

A List of Permissible Section Titles

- introduction
- abstract
- method
- methods
- results
- discussion
- discussions
- conclusion
- conclusions
- results and discussion
- related work
- experimental results
- literature review
- experiments
- background
- methodology
- conclusions and future work
- related works
- limitations
- procedure
- material and methods
- discussion and conclusion
- implementation
- evaluation
- performance evaluation
- experiments and results
- overview
- experimental design
- discussion and conclusions
- results and discussions
- motivation
- proposed method
- analysis
- future work
- results and analysis
- implementation details

B List of Regular Expressions

Citation format regexes:

- `\[([0-9]+\s*[,-;]*\s*)*[0-9]+\s*\]`
- `\(?[12][0-9]3[a-z]?\s*\)`

Hanging citation regex:

```
\s+\(?(\(\s*\)|like|reference|including|include|with|for instance|for example|see also|at|following|of|from|to|in|by|see|as|e\.?g\.?(,)?|viz(\.)?(,)?)\s*(,)*(-)*[\)\]]?\s*[.?!]\s*$
```

C Full Domain Analysis

The full results for all domain adaptation settings can be found in Table 7.

D Reproducibility

D.1 Computing Infrastructure

All experiments were run on a shared cluster. Requested jobs consisted of 16GB of RAM and 4 Intel Xeon Silver 4110 CPUs. We used a single NVIDIA Titan X GPU with 12GB of RAM.

D.2 Average Runtimes

The average runtime performance of each model is given in Table 8. Note that different runs may have been placed on different nodes within a shared cluster.

D.3 Number of Parameters per Model

The number of parameters in each model is given in Table 9.

D.4 Validation Performance

The validation performance of each tested model is given in Table 10.

D.5 Evaluation Metrics

The primary evaluation metric used was F1 score. We used the sklearn implementation of `precision_recall_fscore_support` for F1 score, which can be found here: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html. Briefly:

$$p = \frac{tp}{tp + fp}$$

$$r = \frac{tp}{tp + fn}$$

$$F1 = \frac{2 * p * r}{p + r}$$

where tp are true positives, fp are false positives, and fn are false negatives.

D.6 Hyperparameters

Logistic Regression We used a C value of 0.1151 for logistic regression.

Test Train	Ch	E	CS	P	B	Ec	MS	Ma	Md	Ph
Ch	**63.52**	52.15	51.56	55.08	64.33	52.01	59.95	56.24	70.61	55.93
E	63.04	<u>55.04</u>	55.14	58.39	64.13	55.86	**60.51**	57.93	70.51	57.03
CS	62.29	**55.06**	**56.16**	58.56	63.81	56.13	59.87	58.06	70.22	57.47
P	62.12	53.21	52.53	**60.32**	<u>65.24</u>	<u>56.26</u>	59.59	<u>58.11</u>	70.83	53.96
B	<u>63.06</u>	52.62	52.41	58.71	**65.56**	54.98	59.89	55.07	<u>71.18</u>	55.39
Ec	62.36	54.37	53.02	<u>59.71</u>	64.68	**57.30**	59.45	56.55	70.30	55.39
MS	62.99	51.94	51.57	53.93	63.46	50.22	<u>60.39</u>	56.98	70.45	56.82
Ma	62.81	54.97	<u>55.36</u>	57.59	64.04	55.20	60.28	**59.98**	69.93	<u>57.87</u>
Md	62.26	52.27	51.34	59.19	64.35	55.44	59.08	54.41	**71.19**	53.59
Ph	61.89	52.85	54.12	52.84	63.07	50.67	59.59	<u>58.11</u>	69.06	**58.83**
σ	0.521	1.277	1.757	2.570	0.758	2.499	0.529	1.870	0.728	1.685

Table 7: F1 performance on different domain adaptation settings for the fields (Ch)emistry, (E)ngineering, (C)omputer (S)cience, (P)sychology, (B)iology, (Ec)onomics, (M)aterials (S)cience, (Ma)thematics, (M)e(d)icine, and (Ph)ysics. Out-of-domain tests use the entire set of data from that field, while in domain tests use 80% of data for training, 10% for validation, and 10% for test. Best training domain in bold, second best is underlined.

Setting	Time
Logistic Regression	00h01m43s
Transformer	02h55m13s
BERT	05h30m30s
SciBERT (no weighting)	09h22m00s
SciBERT	09h32m37s
SciBERT + PU	16h01m27s

Table 8: Average runtimes for each model (runtimes are taken for the entire run of an experiment).

Method	# Parameters
Logistic Regression	198,323
Transformer	9,789,042
BERT	109,484,290
SciBERT	109,920,514

Table 9: Number of parameters in each model

Method	F1
Logistic Regression	-
Transformer	57.02
BERT	60.75
SciBERT (no weighting)	57.52
SciBERT	62.04
SciBERT + PU	61.43

Table 10: Average validation performance for each of the models.

Basic Transformer The final hyperparameters for the basic transformer model are: batch size: 64; number of epochs: 33; feed-forward dimension: 128; learning rate: 0.0001406; number of heads: 3; number of layers: 5; weight decay: 0.1; dropout probability: 0.4. We performed a Bayesian grid search over the following ranges of values, optimizing validation F1 performance: learning rate: $[0.000001, 0.001]$; batch size: $\{4, 8, 16, 32, 64, 128\}$; weight decay: $\{0.0, 0.0001, 0.001, 0.01, 0.1\}$; dropout probability: $\{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$; number of epochs: $[2, 40]$; feed-forward dimension: $\{128, 256, 512, 1024, 2048\}$; number of heads: $\{1, 2, 3, 4, 5, 6, 10, 12\}$; number of layers: $[1, 12]$.

BERT The final hyperparameters for BERT are: batch size: 8; number of epochs: 3; learning rate: 0.000008075; triangular learning rate warmup steps: 300; weight decay: 0.1; dropout probability: 0.1. We performed a Bayesian grid search over the following ranges of values, optimizer validation F1 performance: learning rate: $[0.0000001, 0.0001]$; triangular learning rate warmup steps: $\{0, 100, 200, 300, 400, 500, 1000, 1500, 2000, 2500, 5000\}$; batch size: $\{4, 8\}$; weight decay: $\{0.0, 0.0001, 0.001, 0.01, 0.1\}$; number of epochs: $[2, 40]$.

SciBERT The final hyperparameters for SciBERT are: batch size: 4; number of epochs: 3;

learning rate: 0.000001351; triangular learning rate warmup steps: 300; weight decay: 0.1; dropout probability: 0.1. We performed a Bayesian grid search over the following ranges of values, optimizer validation F1 performance: learning rate: $[0.0000001, 0.0001]$; triangular learning rate warmup steps: $\{0, 100, 200, 300, 400, 500, 1000, 1500, 2000, 2500, 5000\}$; batch size: $\{4, 8\}$; weight decay: $\{0.0, 0.0001, 0.001, 0.01, 0.1\}$; number of epochs: $[2, 40]$.

D.7 Data

CITEWORTH is constructed from the S2ORC dataset, which can be found here: `https://github.com/allenai/s2orc`. In particular, CITEWORTH is built using the `20200705v1` release of the data. We plan to release CITEWORTH upon acceptance of the paper, and will then include a link describing how and where to acquire the data.