CITEWORTH: Cite-Worthiness Detection for Improved Scientific Document Understanding

Dustin Wright and **Isabelle Augenstein**
Dept. of Computer Science
University of Copenhagen
Denmark
{dw|augenstein}@di.ku.dk

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 many potential training signals, such as citations, which can be used to build large labelled datasets. Given this, we present an in-depth study of cite-worthiness detection in English, where a sentence is labelled for whether or not it cites an external source. To accomplish this, we introduce CITEWORTH, a large, contextualized, rigorously cleaned labelled dataset for cite-worthiness detection built from a massive corpus of extracted plain-text scientific documents. We show that CITEWORTH is high-quality, challenging, and suitable for studying problems such as domain adaptation. Our best performing cite-worthiness detection model is a paragraph-level contextualized sentence labelling model based on Longformer, exhibiting a 5 F1 point improvement over SciBERT which considers only individual sentences. Finally, we demonstrate that language model fine-tuning with cite-worthiness as a secondary task leads to improved performance on downstream scientific document understanding tasks.

1 Introduction

Building effective NLP 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 extracted, structured, and unlabelled plain-text scientific documents 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 systems for scientific document processing?

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). 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 poorly removed citation text causing ungrammatical sentences and hanging punctuation, which can trivially indicate a cite-worthy or non-cite-worthy sentence. Additionally, the task itself is quite difficult to learn, as different fields employ citations differently, and whether or not a sentence contains a citation depends on factors such as the context in which it

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1796–1807

appears. Given this, we present CITEWORTH, a rigorously curated dataset for cite-worthiness detection in English. CITEWORTH contains rich metadata, such as authors and links to cited papers, and all data is provided in *full paragraphs*: every sentence in a paragraph is labelled in order to provide sentence *context*. We offer the dataset to the research community to facilitate further research on cite-worthiness detection and related scientific document processing tasks.

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**: What methods are most effective for automatically detecting cite-worthy sentences (§4)?
>
> **RQ3**: How does domain affect learning cite-worthiness detection (§5)?
>
> **RQ4**: 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, that there are large differences in how models generalize to data from different fields, and that sentence context leads to significant performance improvements on cite-worthiness detection. 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 method for cite-worthiness detection which considers the entire paragraph a sentence resides in, improving by 5 F1 points over the state of the art model for scientific document processing, SciBERT (Beltagy et al., 2019).
- A thorough analysis of the problem of cite-worthiness detection, including explanations of predictions and insight into how scientific domain affects performance.
- New state of the art on citation intent detection via transfer learning from joint citation detection and language model fine-tuning on CITEWORTH, 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 of the datasets is performed to understand the quality of the data or what aspects of the problem are difficult or easy. Additionally, no study to date has considered how sentence context can affect learning to perform cite-worthiness detection.

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;

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

Luan et al., 2018) and linking (Wright et al., 2019), keyphrase extraction (Augenstein et al., 2017; Augenstein and Søgaard, 2017), relation extraction (Kringelum et al., 2016; Luan et al., 2018), dependency parsing (Kim et al., 2003), citation prediction (Holm et al., 2020), citation intent classification (Jürgens et al., 2018; Cohan et al., 2019), summarization (Collins et al., 2017), 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 automatically curated with low noise? To answer this, we start with the S2ORC dataset of extracted plaintext scientific articles (Lo et al., 2020). It consists of data from 81.1M English scientific articles, with full structured text for 8.1M articles. S2ORC uses SCIENCEPARSE[2] to parse PDF documents and GROBID[3] to extract structured data from text. As such, the data also includes rich metadata, e.g. Microsoft Academic Graph (MAG) categories, linked citations, and linked figures and tables. Throughout this work, a "citation span" denotes a span containing citation text (e.g. "[2]"), and a "citation marker" is any text that trivially indicates a citation, such as the phrase "is shown in." A citation span is also a type of citation marker. It is important to remove all citation markers from the dataset to prevent the model learning to use these signals for prediction.

3.1 Data Filtering

Given the size of S2ORC, we first reduce the candidate set of data to papers where all of the following are available.

- Abstract
- Body text
- Bibliography
- Tables and figures
- Venue information
- Inbound citations
- Microsoft Academic Graph categories

Filtering based on these criteria results in 5,494,387 candidate papers from which to construct the dataset. After filtering the candidate set of papers, we perform the following checks on the sentences in the body text.

1. Citation spans are parenthetical author-year or bracketed-numerical form.
2. Citation spans are at the end of a sentence.
3. All possible citation spans have been extracted by S2ORC.
4. No citation markers are left behind after removing citation spans from the text.
5. Sentence starts with a capital letter, ends with '.', '!', or '?', and is at least 20 characters long.

The detailed steps of extracting and labelling sentences based on these criteria are given in §3.2. With the first two criteria, we restrict 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 constrained to those of the following forms.

> This result has been shown in previous work (Author1 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", as well as any sentence with a citation format that does not match the two we have selected.

Curating 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 (for example, the above sample would turn into "The work *of has* shown this in previous work"), 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.

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

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

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.

3.2 Extracting Cite-Worthy Sentences in Context

As we are interested in using sentence context for prediction, we perform extraction at the *paragraph level*, ensuring that all of the sentences in a given paragraph meet the checks given in §3.1. As such, our dataset construction pipeline for a given paper begins by first extracting all paragraphs from the body text which belong to sections with titles coming from a constrained list of permissible titles (e.g. "Introduction," "Methods," "Discussion") . The full list is provided in Appendix A.

For a given paragraph, we first word and sentence tokenize the text with 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 (see Appendix B). If a match is found the paragraph is ignored, as we only consider paragraphs where all citations have been extracted by S2ORC. Otherwise, the location and format of the citation is checked, again using regular expressions (see Appendix B). If the citation is not at the end of the sentence, the paragraph is ignored. We then remove the citation text using the provided citation spans for all sentences which pass the above checks.

Simply removing the citation span runs the risk of leaving other types of citation markers, such as 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. The regular expression checks for many common prepositional phrases and citation markers occurring as the last phrase of a sentence such as "see," "of," "by," etc.

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 (31.76%)
Total non-cite-worthy	806,405 (68.24%)
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.

To handle issues with sentence tokenization, 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, the paragraph is 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.[4]

3.3 Manual Evaluation

In order to provide some measure of the general quality of CITEWORTH, we perform a manual eval-

[4]The full dataset can be downloaded from this repository: https://github.com/copenlu/cite-worth

Method	Extracted Correct	Markers Removed
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 . "Extracted Correct" are results for correctly extracting the sentences (i.e. that sentences are tokenized correctly and are grammatical), and "Markers Removed" are results for successfully removing citation markers. The data curated using our method has 6% fewer errors in terms of extraction and removal of citation markers, and less than 2% of the samples have some form of citation marker.

uation 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 extraction 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. 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 expert annotator with a PhD in computer science for labelling. The annotator is instructed to label if the sentences are complete and have no hanging punctuation or obvious extraction errors, and if there are any textual indicators that the sentences contain a citation. The results for the manual annotation 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 extraction 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 sentence cleaning 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 supervised learning.

4 RQ2: System Evaluation[5]

Next, we ask: what methods are most effective for performing cite-worthiness detection? To answer this and characterize the difficulty of the problem, we run a variety of baseline models on CITEWORTH. The hyperparameters selected for each model, as well as hyperparameter sweep information, are given in Appendix C.6.

Logistic Regression As a simple baseline, we use a logistic regression model with TF-IDF input features.

Färber et al. (2018b) The convolutional recurrent neural network (CRNN) model from Färber et al. (2018b). They additionally use oversampling to deal with class imbalance.

Transformer We additionally train a 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.

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 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). The intuition behind PU learning is to assume that cite-worthy data is labelled and non-cite-worthy data is unlabelled, containing some cite-worthy examples. This is to mitigate the subjectivity involved in adding citations to sentences. Technically, this involves training a classifier on the positive-unlabeled data which will predict the probability that a sample is labeled, and using this to estimate the probability that a sample is positive given that it is unlabeled. One then trains a second model where positive samples are trained on normally and unlabeled samples are duplicated and trained on twice, once as positive and once as negative data, weighed by the first model's estimate of the probability that the sample is positive.

Longformer-Ctx Finally, we test our novel contextualized prediction model based on Longformer (Beltagy et al., 2020). Longformer is a

[5]The code for all experiments can be found here: https://github.com/copenlu/cite-worth

Method	P	R	F1
Logistic Regression	$46.65_{0.00}$	$64.88_{0.00}$	$54.28_{0.00}$
Färber et al. (2018b)	$49.57_{0.96}$	$65.56_{2.61}$	$56.41_{0.34}$
Transformer	$47.92_{0.78}$	$71.59_{1.74}$	$57.39_{0.10}$
BERT	$55.04_{0.66}$	$69.02_{1.33}$	$61.23_{0.21}$
SciBERT-no-weight	$\mathbf{65.94}_{0.37}$	$51.62_{0.53}$	$57.91_{0.30}$
SciBERT	$57.03_{0.50}$	$68.08_{1.03}$	$62.06_{0.15}$
SciBERT + PU	$49.46_{0.83}$	$\mathbf{82.12}_{1.40}$	$61.73_{0.27}$
Longformer-Solo	$57.21_{0.25}$	$68.00_{0.41}$	$62.14_{0.02}$
Longformer-Ctx	$59.92_{0.28}$	$77.15_{0.49}$	$\mathbf{67.45}_{0.06}$

Table 4: F1 performance of baselines on the test set of CITEWORTH. Results are averaged across 5 seeds, with standard deviations given in the subscripts.

Transformer based language model which uses a sparse attention mechanism to scale better to longer documents. We process an entire paragraph at a time, separating each sentence with a `[SEP]` token. Each `[SEP]` token representation at the output of Longformer is then passed through a network with one hidden layer and a classifier. As a control, we also experiment with Longformer using only single sentences as input (Longformer-Solo).

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.

Our results indicate that context is critical, resulting in the best F1 score of 67.45 (Longformer-Ctx) and a 5.31 point improvement over the next best model. Using class weighting is also highly important, resulting in another increase of over 4 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 BERT, SciBERT, and Longformer all perform significantly better than a Transformer trained from scratch and the model from Färber et al. (2018b).

To gain some insight into what the model learns, we visualize the most salient features from SciBERT for selected easy and hard examples. We use the single-sentence model instead of the paragraph model for simplicity. "Easy" samples are defined as those which the model predicted correctly with high confidence, and "hard" examples are defined as those for which the model had low confidence in its prediction. 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 method works by calculating the gradient of the output with respect to the input, then multiplies this with the input. In the examples below "C" refers to an example whose gold label is cite-worthy, and "N" refers to an example whose gold label is non-cite-worthy.

The model is able to pick up on obvious markers of cite-worthy and non-cite-worthy sentences for the following correctly classified examples, such as that a sentence refers to a preprint or to different sections within the paper itself:

C: [CLS] in this note , we follow the approach to the en ##och ##s conjecture outlined in the preprint . [SEP]

N: [CLS] conclusions are provided in section 4 . [SEP]

We also see that the dataset contains many relatively difficult instances, as we show in the following incorrectly classified examples. E.g., the model observes "briefly discussed" as an indicator that an instance is non-cite-worthy when it is in fact cite-worthy, and that "described earlier" and "previous work" signal that a sentence is cite-worthy when it is in fact labelled as non-cite-worthy.

C: [CLS] some approaches for the solution as well as their limitations are briefly discussed . [SEP]

N: [CLS] this simple and fast technique for the production of snps was described earlier in our previous work . [SEP]

We hypothesize that in such instances, context can help the most in disambiguating which sentences in a paragraph should be labelled as cite-worthy. Additionally, other information such as the section in which a sentence resides could help. E.g., to correctly label the fourth statement above as "non-cite-worthy", it may help to see that the last sentence of the paragraph is "In our previously published work, it was reported that SNPs were joined together by the heat treatment, and this process led to increase in the sizes of SNPs which finally resulted in sharper XRD peaks" which is a cite-worthy sentence. Additionally, it may help to know that it resides in the "Discussion" section of the paper.

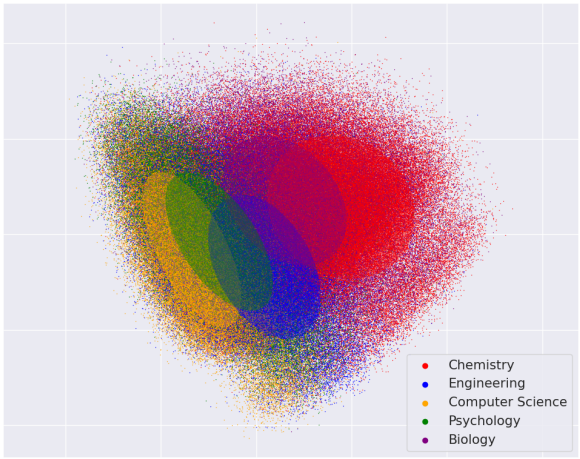

Figure 1: Visualizing the BERT embeddings for 5 of the 10 domains from CITEWORTH using the method by Aharoni and Goldberg (2020). Clustering is performed using Gaussian Mixture Models.

Train \ Test	Ch	E	CS	P	B
Ch	**67.58**	58.41	56.86	62.35	68.23
E	66.62	**60.25**	60.11	64.02	68.07
CS	65.05	59.36	**61.99**	63.85	66.72
P	65.49	58.03	56.69	**65.10**	68.27
B	66.59	58.80	58.22	64.54	**69.12**
σ	0.90	0.78	2.02	0.92	0.77
ρ	0.87	0.86	0.76	0.67	0.79

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. σ is the standard deviation of performance of different train domains on the given test domain, and ρ is Pearson correlation between performance and Euclidean distance from the train domain cluster to the test domain cluster.

5 RQ3: Domain Evaluation

We next ask: how does domain affect learning to perform cite-worthiness? To answer this, we study the relationships between cite-worthiness data from different fields and how the Longformer-Ctx model 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: Chemistry, Engineering, Computer Science, Psychology, and Biology.

First, 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 (specifically the base, uncased variant) and the output representations for each token in a sentence are average pooled. These representations are visualized in 2D space via PCA in Figure 1. 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, thus differences could exist in a model's ability to perform cite-worthiness detection on out of domain data.

To test this, 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 Longformer-Ctx are given in Table 5.

Not surprisingly, the best performance for each split occurs when training on data from the same field. We also observe high variance in the maximum performance for each field (σ = 3.32), and between different fields on the same test data, despite large pretrained Transformer models being relatively invariant across domains (Wright and Augenstein, 2020b). This suggests stark differences in how different fields employ citations. Additionally, we observe a strong (inverse) correlation between distance in the embedding space and performance on different domains, showing that using more similar data for training helps on out-of-domain performance (Aharoni and Goldberg, 2020).

6 RQ4: 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. We use SciBERT in order to have a direct comparison with previous work (Beltagy et al., 2019). The tasks we evaluate on come from Beltagy et al. (2019) and are categorized as follows.

Dataset	Reference	Task	Base	LM	Cite	LMCite
BC5CDR	Li et al. (2016)	NER	$89.84_{0.18}$	$\mathbf{90.03_{0.11}}$	$89.73_{0.25}$	$90.02_{0.79}$
JNLPBA	Kim et al. (2004)	NER	$77.02_{0.36}$	$77.13_{0.53}$	$76.97_{0.44}$	$\mathbf{77.15_{0.58}}$
NCBI-Disease	Doğan et al. (2014)	NER	$\mathbf{88.79_{0.35}}$	$88.53_{0.58}$	$88.66_{0.57}$	$88.31_{0.43}$
SciERC	Luan et al. (2018)	NER	$67.08_{0.50}$	$66.64_{0.47}$	$67.12_{0.46}$	$\mathbf{67.48_{0.45}}$
EBM-NLP	Nye et al. (2018)	PICO	$76.61_{0.21}$	$\mathbf{76.69_{0.28}}$	$76.55_{0.88}$	$76.41_{0.32}$
ChemProt	Kringelum et al. (2016)	REL	$83.17_{0.43}$	$\mathbf{83.26_{0.90}}$	$82.70_{1.06}$	$83.16_{0.63}$
SciERC	Luan et al. (2018)	REL	$80.21_{0.81}$	$\mathbf{80.68_{1.04}}$	$80.00_{1.73}$	$80.58_{0.96}$
ACL-ARC	Jürgens et al. (2018)	CLS	$71.82_{2.93}$	$70.95_{2.25}$	$\mathbf{73.68_{2.75}}$	$72.92_{3.76}$
SciCite	Cohan et al. (2019)	CLS	$84.83_{0.65}$	$85.18_{0.47}$	$85.32_{0.16}$	$\mathbf{85.35_{0.29}}$
PaperField	Beltagy et al. (2019)	CLS	$65.48_{0.18}$	$\mathbf{65.57_{0.27}}$	$65.46_{0.24}$	$65.42_{0.48}$
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 the same as in their paper: 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. Subscripts are the standard deviation for 5 runs.

- Named Entity Recognition (NER)/PICO: These tasks involve labelling the spans of different types of entities in a document.
- Relation Extraction (REL): This task involves labelling a sequence for the relationship between two entities.
- Text classification (CLS): Finally, we test on several text classification tasks (citation intent classification and paper field classification), where the goal is to classify a sentence into one or more categories.

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

Base The original SciBERT model.

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.

LMCite 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 of 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.[6]

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. This is in line with previous work reporting that language model fine-tuning on in-domain data leads to improvements on end-task fine-tuning (Gururangan et al., 2020). 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; present a

[6]We release two CITEBERT models available from the HuggingFace model hub: `copenlu/citebert` and `copenlu/citebert-cite-only`.

paragraph-level contextualized model which improves by 5.31 F1 points on the task of cite-worthiness detection over the existing state-of-the-art; 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, one could explore other useful problems such as modeling different authors' writing styles and incorporating the author network as a signal. We hope that the data and accompanying fine-tuned models will be useful to the research community working on problems in the space of scientific language processing.

Acknowledgements


The research documented in this paper has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801199.

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.

Isabelle Augenstein and Anders Søgaard. 2017. Multi-task learning of keyphrase boundary classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 341–346, Vancouver, Canada. 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 3606–3611.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *CoRR*, abs/2004.05150.

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.

Ed Collins, Isabelle Augenstein, and Sebastian Riedel. 2017. A supervised approach to extractive summarisation of scientific papers. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 195–205, Vancouver, Canada. Association for Computational Linguistics.

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

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.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics.

Andreas Nugaard Holm, Barbara Plank, Dustin Wright, and Isabelle Augenstein. 2020. Longitudinal Citation Prediction using Temporal Graph Neural Networks. *arXiv preprint arXiv:2012.05742*.

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.

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.

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.

Dustin Wright, Yannis Katsis, Raghav Mehta, and Chun-Nan Hsu. 2019. NormCo: Deep Disease Normalization for Biomedical Knowledge Base Construction. In *AKBC*.

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*$`

Setting	Time
Logistic Regression	00h01m43s
Transformer	02h55m13s
BERT	05h30m30s
SciBERT (no weighting)	09h22m00s
SciBERT	09h32m37s
SciBERT + PU	16h01m27s
Longformer-Solo	75h27m22s
Longformer-Ctx	19h16m07s

Table 7: 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
Longformer	149,251,586

Table 8: Number of parameters in each model

C Reproducibility

C.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.

C.2 Average Runtimes

The average runtime performance of each model is given in Table 7. Note that different runs may have been placed on different nodes within a shared cluster.

C.3 Number of Parameters per Model

The number of parameters in each model is given in Table 8.

C.4 Validation Performance

The validation performance of each tested model is given in Table 9.

Method	F1
Logistic Regression	-
Transformer	57.02
BERT	60.75
SciBERT (no weighting)	57.52
SciBERT	62.04
SciBERT + PU	61.43
Longformer-Solo	61.67
Longformer-Ctx	67.11

Table 9: Average validation performance for each of the models.

C.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.

C.6 Hyperparameters

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

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: $[2, 40]$.

Longformer-Ctx The final hyperparameters for Longformer-Ctx are: batch size: 4; number of epochs: 3; learning rate: 0.00001112; triangular learning rate warmup steps: 300; weight decay: 0.0; 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: $[2, 6]$.

C.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. A link to the CITEWORTH data can be found here: https://github.com/copenlu/cite-worth.