Improving State of the Art in Context-Aware Citation Recommendation

Authors
Ciaran Coleman - 909473
Ahmed Adeeb Fawzy - 17090863
Steven McDonald - 20074556
Anjana Ratnayake - 17016257

Abstract

Context-aware citation recommendation aims to recommend suitable reference citations when provided with the text adjacent to a citation placeholder. (Chanwoo Jeong, 2019) achieves state of the art results when combining a pre-trained model for context encoding – Bidirectional Encoder Representations from Transformers (BERT) (J. Devlin and Toutanova, 2018) – and a Graph Convolution Network (GCN) (Kipf and Welling, 2016) for metadata encoding.

We use this as a foundation for our work, reproducing some of their key results and proposing further developments. We show that encoding the citation context with a domain-specific BERT variant SciBERT achieves a 14% increase in mean average precision (MAP). We are not able to reproduce the uplift seen through encoding metadata with a GCN and test alternative metadata encoding using a feature-based approach, again with SciBERT.

1 Introduction

Citations play a fundamental role in academic writing. Citing is not only important for crediting others for prior work, but also to allow efficient validation of scientific claims and facilitates communication between author and reader (Michael Färber, 2020).

The traditional approach of keyword searches through academic literature databases requires heavy involvement of the researcher and is steadily becoming prohibitive due to exponential growth in scientific publications within certain domains (He et al., 2011). Research into automated citation recommendation, the task of recommending suitable publications for use as citations within an input document (Michael Färber, 2020), has recently gained traction to tackle this issue. Coupled with the resurgence of deep learning in the past decade (Goodfellow et al., 2016), model architectures based on neural networks – in particular, that of pre-trained language models such as Bidirectional Encoder Representations from Transformers (BERT) (J. Devlin and Toutanova, 2018) – have shown promise for this task.

The work of (Chanwoo Jeong, 2019) is one example that achieves state of the art (SOTA) results for local citation recommendation on their own benchmark FullTextPeerRead dataset. To further their performance, we hypothesise the following:

H1 Fine-tuning a domain-adapted BERT model as citation context encoder will lead to new SOTA on the dataset, over BERT-GCN.

H2 Adding a domain-adapted BERT model as metadata encoder, SciBERT-META, can enrich the citation context embedding and lead to further improvements in performance.

H3 Naively applying a fixed context length leads to incomplete words that could give misinformed representations of the context. Therefore removing partial words from the context could lead to improved performance, even with slightly reduced context length.

The remainder of this paper is organised as follows: in Section 2 we briefly discuss BERT and its variants, as well as recent, neural network-based approaches to context-aware citation recommendation. Section 3 introduces the dataset and discloses any peculiarities the reader should be made aware of, before developing our approach to modelling, training and evaluation in Section 4. Experiments to improve and understand performance changes from our proposals are detailed in Section 5, with a discussion of results in Section 6. Our findings are then concluded in Section 7.

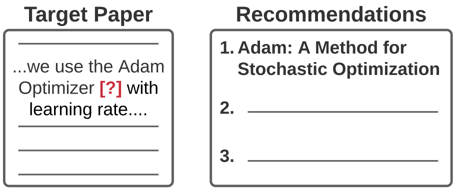

Target Paper
...we use the Adam Optimizer [?] with learning rate....
Recommendations
1. Adam: A Method for Stochastic Optimization
2.
3.

Figure 1: An example context-aware citation recommendation. The placeholder is shown in red.

2 Related Work

We introduce the concept of citation recommendation and focus the discussion on works that involve neural architectures, as these are generally the most recent and relevant. However, we encourage the reader to refer to (Michael Färber, 2020) for a comprehensive overview on other approaches.

2.1 Citation Recommendation

At a high level, citation recommendation can be tackled either globally or locally. With global citation recommendation, the entire document – or perhaps just title and abstract – are used as input to generate a list of candidate references. Although straightforward, it does not address the need to place the citations at appropriate locations in the manuscript. Local, or context-aware, citation recommendation (He et al., 2010) solves this by recommending citations for specific claims or fragments of text adjacent to a citation placeholder (the citation context). Figure 1 shows an example of this. We consider only context-aware citation recommendation as it more naturally follows the scientific writing process.

In addition to citation context, metadata such as author, title, journal/conference, etc. are typically available. Recent approaches have shown that they can aid in the task through enrichment of citation context representations. For example, (Ebesu and Fang, 2017) propose an encoder-decoder architecture called Neural Citation Network (NCN) that generates robust representations of citation context, further augmented by embeddings of the cited paper's author(s). Concatenation of the two allows the interaction between context and author to be learned during the decoding process.

(Dai, 2018) also explores neural architectures enriched by metadata, arguing for representing papers with only author and venue information. Separate representations of the citation contexts in scientific papers are learned using Long Short-Term Memory (Hochreiter and Schmidhuber, 1997) networks, before computing relevance scores between them that for ranking purposes.

(Chanwoo Jeong, 2019) consider instead to use a pre-trained language model (BERT) as a citation context encoder. To enrich these embeddings, relationship embeddings between target (citing) and source (cited) papers are learned by training a GCN using available paper metadata. The authors report that these relationship embeddings improve performance. This improvement however, is marginal.

2.2 Bidirectional Encoder Representations from Transformers

Pre-trained language models have come to form the backbone of many natural language processing tasks. In recent years, language models that embed words based on context such as ELMo (Peters et al., 2018) and BERT, have steadily claimed SOTA over context-free models such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). BERT makes use of a transformer architecture (Vaswani et al., 2017) to be able to read all the context simultaneously and is trained as a Masked Language Model [1]. One advantage of BERT is that it uses subword tokenisation (Schuster and Nakajima, 2012), allowing ease of handling out of vocabulary (OOV) words with a small footprint.

BERT Variants

Since the success of BERT, considerable effort has been placed in adapting the model architecture to different domains, ranging from scientific texts, to patent texts (Lee and Hsiang, 2019) and even to video data (Sun et al., 2019).

Variants typically involve either fine-tuning or pre-training BERT on the domain-specific corpora. Pre-trained adaptations have generally shown better empirical performance (e.g. (Iz Beltagy, 2019)). For the task of citation recommendation, the following pre-trained variants are deemed well-suited. In both cases, the architecture and training process are consistent with original BERT model.

1. SCIBERT (Iz Beltagy, 2019): Two key points separate this model from BERT. Firstly, SciBERT is trained exclusively on a corpus of scientific publications, consisting of 18% computer science papers with the remaining 82% in the biomedical field. Second is the construction of a new vocabulary, SciVocab, to reflect the scientific corpus.

[1] see (J. Devlin and Toutanova, 2018) for more details

Table 1: FullTextPeerRead reported and released details

Detail	Reported	Released
# of total papers	4,898	4,837
# of base papers	3,761	3,695
# of cited papers	2,478	2,444
# of citation context	17,247	16669
years of published papers	2007-2017	2007-2017

2. BIOBERT: (Lee et al., 2019): BioBERT is pretrained on a mixture of general corpora and biomedical corpora of PubMed abstracts and PubMedCentral full-text articles. In contrast to SciBERT, BioBERT maintains the same vocabulary as the original BERT. There are only cased versions.

3 Datasets

Several datasets are available with (Michael Färber, 2020) providing a recent overview, but finding one that has all the right features can be difficult. For example, CiteSeerX (et al., 2014) boasts over 2M indexed documents, but is known to be noisy and lacking metadata (Chanwoo Jeong, 2019). The ACL-AAN dataset (Dragomir R. Radev and Abu-Jbara, 2013) is significantly smaller than CiteSeerX with c. 25K indexed papers[2] from the field of computational linguistics. Unfortunately, neither citation context nor metadata are provided directly, instead requiring extraction from a database.

3.1 FullTextPeerRead

Two new datasets are proposed in (Chanwoo Jeong, 2019), one of which is unavailable due to a policy prohibiting disclosure of modified data.

The second, FullTextPeerRead, is a modified version of the PeerRead dataset (Dongyeop Kang, 2018). PeerRead itself consists of over 14K paper drafts from the Computer Science domain and includes other information such as accept/reject decisions and peer reviews for a subset. FullTextPeerRead re-formats the dataset to the task of context-aware citation recommendation. We find slight differences in the size of the released dataset compared to that originally reported in (Chanwoo Jeong, 2019) (Table 1).

Each entry in FullTextPeerRead includes one citation text to the left and one to the right of the citation placeholder. The following complete metadata are also provided for both target and source paper: paper id, title, abstract, authors, venue and year of publication.

Manual Assessment of FullTextPeerRead

(Chanwoo Jeong, 2019) mentions manual post-processing to remove noisy data in FullTextPeerRead, citing inconsistent formatting of the manuscripts as the reasoning. As full details are not provided, we conduct a manual review to check whether any noise remains. We randomly sampled a small set of 50 citation contexts from the complete dataset, each with a maximum sequence length of 50, and review for irregularities. We find two frequent noise sources that could affect performance:

1. Adjacent words mistakenly joined together. This was found in $\approx 23\%$ of citation contexts, and could adversely affect the overall semantic meaning of a short citation context. An example found is *'textreconstruction'*. When the words are joined, it is tokenized as ['text', '##rec', '##onstr','##uct', '##ion'] instead of ['text', 'reconstruction'] when the words are separated.
2. Incomplete words within the body of context. This appeared in $\approx 17\%$ of the samples and it is generally the case that the beginning of the word is missing. For example, 'Bayesian' is instead 'yesian' which could lead to a very different representation of the context.

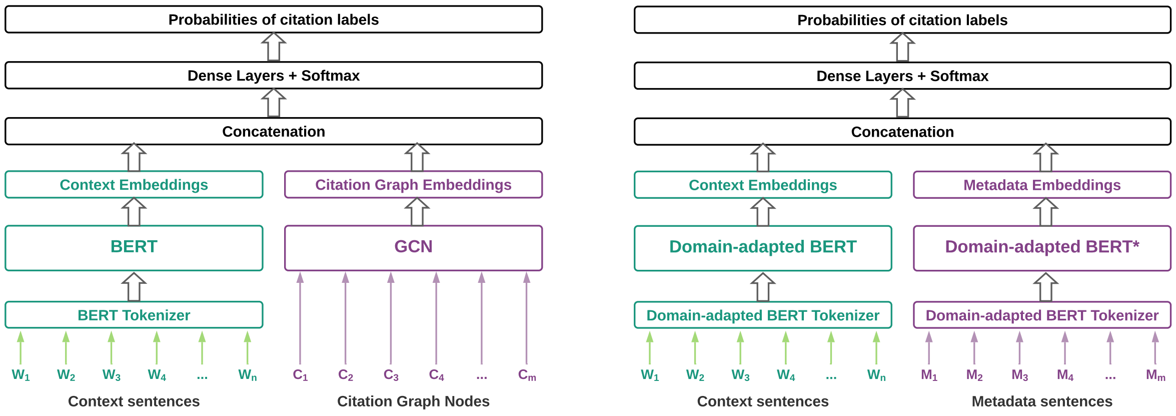
Both identified sources are difficult to correct for without manual input. We elect to retain these examples to enable direct comparison of results between works.

We also review the presence of any imbalances in the dataset. In particular, we note a class imbalace whereby a small number of source papers are cited much more often than the majority. Figure 3 shows that 77.2% of source papers are only cited between 1 and 5 times in the dataset, but the range extends up to 572 citations. 15 papers, representing 0.6% of all source papers, contribute to over 22.2% of all citations in the dataset. This imbalance could lead to models showing preference to often-cited papers (i.e. overfit); this may, however, mimic what occurs in reality where particular seminal papers see a disproportionate amount of citations.

4 Methods

In this section we describe our modelling approach to improve upon the BERT-GCN in (Chanwoo Jeong, 2019), as well as detailing the training and evaluation procedure.

[2] 24,622 at the time of writing, http://aan.how


Probabilities of citation labels
Dense Layers + Softmax
Concatenation
Context Embeddings
Citation Graph Embeddings
BERT
GCN
BERT Tokenizer
W_1 W_2 W_3 W_4 ... W_n
C_1 C_2 C_3 C_4 ... C_m
Context sentences
Citation Graph Nodes
Probabilities of citation labels
Dense Layers + Softmax
Concatenation
Context Embeddings
Metadata Embeddings
Domain-adapted BERT
Domain-adapted BERT*
Domain-adapted BERT Tokenizer
Domain-adapted BERT Tokenizer
W_1 W_2 W_3 W_4 ... W_n
M_1 M_2 M_3 M_4 ... M_m
Context sentences
Metadata sentences

(a) **(b)**

Figure 2: Comparison of model architectures: a) Original BERT-GCN model proposed by (Chanwoo Jeong, 2019) and b) Our proposed model. *The model parameters that generate the metadata embeddings are frozen.

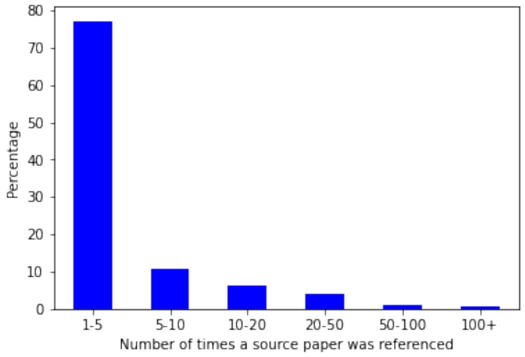


Figure 3: Breakdown of the proportion of source papers for different citation frequencies within the dataset.

4.1 Model Overview: XXXBERT-META

We begin by outlining two proposed changes to the BERT-GCN model, motivated by the marginal performance improvement of the GCN in (Chanwoo Jeong, 2019). Our model is compared to theirs in Figure 2. Our model contains a citation context encoder that outputs a citation context embedding from an input tokenized by WordPiece, and a metadata encoder that extracts a textual embedding of the selected metadata, also tokenized by WordPiece. Both encoders have the same base BERT architecture with a hidden size of 768, 12 transformer blocks and 12 attention heads. The context embedding is the pooled output of the [CLS] (classifier) token, while the metadata embeddings can be constructed from other hidden layers of the transformer architecture. These embeddings are concatenated before passing through a single dense layer with softmax normalization to return a probability distribution across citation labels, i.e. the list of potential source papers. A source paper's ranking for retrieval can then be inferred from its probability ranking.

4.2 Citation Context Encoder

The first change we propose is to replace BERT as the context encoder with a variant pre-trained on scientific corpora, as this is the nature of our dataset. These variants are expected to learn representations of the citation context that improve the downstream task.

Table 2: Comparing BERT and SciBERT tokenization

Model	Output
Original BERT	’we optimize the stochastic gradient’ [’we’, ’opt’, ’##imi’, ’##ze’, ’the’, ’st’, ’##och’, ’##astic’, ’gradient’]
SciBERT	[’we’, ’optimize’, ’the’, ’stochastic’, ’gradient’]

We review how each model's tokenization handles a sample context sentence to gain intuition. The example in Table 2 compares SciBERT to BERT to emphasise the impact of different model vocabulary. BERT's tokenizer splits the words ‘optimize’ and ‘stochastic’, commonly encountered in the Computer Science domain, into multiple subwords as these are not common within the general corpora it was trained on; whilst SciBERT's tokenizer handles it perfectly.

4.3 Metadata Encoder

Our second proposal is to replace the GCN with a metadata encoder, also in the form of a domain-adapted variant of BERT, to augment the citation context embedding.

It is demonstrated in (J. Devlin and Toutanova, 2018) that a feature-based approach – applying various aggregation strategies to the hidden layer activations – to create contextualised embeddings, could yield similar performance to a fine-tuned approach[3]. We postulate that applying these methods to the paper metadata could generate semantically meaningful embeddings that may help with the downstream task.

Layer Choice for Representation

Each encoder layer of the BERT architecture can be interpreted as capturing a unique feature representation with different degrees of semantic and syntactic information (Sun et al., 2019).

We take inspiration from (J. Devlin and Toutanova, 2018) on how to construct features, with further details in Section 5. However, we opt to take the average of all token embeddings in the sequence instead of only the embedding at the [CLS] token. In (Reimers and Gurevych, 2019), this is found to be the better of the two approaches when evaluating semantic textual similarity.

Combining Metadata

The title and abstract together provide a summary of the overall content of a manuscript and could be used to enrich the citation context.

To combine metadata, we elect to concatenate the separate text before feeding in to the encoder to produce a single embedding. An analysis of the number of tokens when concatenating title and abstract shows that all but one example are well within the maximum token sequence length of 512.

4.4 Model Training Procedure

The procedures stated here are used as default unless stated otherwise in Section 5.

As this work does not focus on hyperparameter tuning for performance improvement, we follow those used in (Chanwoo Jeong, 2019), re-using their source code [4] where possible. This ensures a fair comparison between models. The hyperparameters for fine-tuning are set to 30 epochs, a mini-batch size of 16 and dropout probability of 0.1. The Adam optimiser (Kingma and Ba, 2015) is used with a learning rate $2e-5$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and L2 weight-decay of 0.01.

Training involves minimising a classification objective of cross-entropy loss. Due to per-example calculation, this training objective does not directly relate to the final metrics (Section 4.5) used to evaluate the paper rankings; therefore, optimising for this objective may not be best for improving the downstream task (Lin et al., 2020). During fine-tuning, the weights of the metadata encoder are considered fixed. For main results, we repeat the fine-tuning procedure three times to account for random initialisation.

The dataset is pre-processed to only consider examples where the target paper has a citation frequency > 5 and both left and right contexts are used. There are a total of $C = 489$ potential source papers based on this for the multi-class classification objective. For experiments with no element of parameter tuning, pre-2017 papers are used for the training set with 2017 papers as the test set. For experiments that involve tuning multiple configurations, e.g. metadata combinations, 2017 test papers are removed and instead pre-2016 papers used for training and 2016 papers for development.

4.5 Evaluation Metrics

We will be using the same evaluation metrics as (Chanwoo Jeong, 2019), namely: Recall@k, Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR). A detailed description with accompanying formulae may be found in (Lin et al., 2020). These evaluation metrics tell us how well a model is performing and allows us to compare if any changes made to the model results in a positive outcome.

5 Experiments

Here, we present a number of experiments aimed at improving the model and methods for context-aware citation recommendation. Please refer back to Section 1 where hypotheses are referenced.

Reproducing Results of (Chanwoo Jeong, 2019): Our first aim is to reproduce selected results on the FullTextPeerRead dataset; both to check the code and robustness of previous work, as well as to set baselines for later experiments. The necessary scripts for training and fine-tuning BERT-Only and BERT-GCN models are provided.

Citation Context Encoder: Once baselines are established, we turn our attention to hypothesis **H1**. To test this, we consider both SciBERT (base, uncased) and BioBERT (v1.1, base, cased [5]) as replacements to original BERT.

[3] Specifically for a Named Entity Recognition (NER) task.

[4] available at https://github.com/TeamLab/bert-gcn-for-paper-citation)

[5] No uncased version could be found at the time of writing.

Table 3: Summary of performance metrics for models trained with a maximum context sequence length of 50 characters (total 100 characters) using data with a minimum citation frequency of 5. All our results are reported as an average of three fine-tuning repeats.

Model	MAP	MRR	Recall@5	Recall@10	Recall@30	Recall@50	Recall@80
BERT Jeong et al	0.415	0.415	0.480	0.520	0.593	0.637	0.689
BERT	0.408	0.408	0.476	0.517	0.591	0.641	0.695
SciBERT	**0.467**	**0.467**	**0.546**	**0.596**	**0.681**	**0.727**	**0.774**
SciBERT whole words	0.464	0.464	0.540	0.590	0.677	0.719	0.768
BioBERT	0.416	0.416	0.482	0.521	0.600	0.650	0.701
BERT-GCN Jeong et al	0.418	0.418	0.486	0.529	0.604	0.650	0.699
BERT-GCN	0.403	0.403	0.467	0.508	0.587	0.634	0.685
SciBERT-GCN	0.465	0.465	0.541	0.589	0.680	0.726	0.773
SciBERT-META	0.464	0.464	0.543	0.595	0.678	0.726	0.769

Context Length: We compare different context sequence lengths using both BERT and SciBERT encoding to determine whether there is a difference in sensitivity between the two. (Chanwoo Jeong, 2019) found that increased sequence length was of benefit to the citation recommendation task where citation frequency is set to 1, but noted diminishing returns as context sequence increased above 100. We continue instead with a citation frequency of 5, testing context sequence lengths of 50, 100 and 150.

Context Pre-processing: To test hypothesis **H3**, we compare standard pre-processing of the citation context using a maximum length cut-off, to a method that then discards incomplete words. The maximum context sequence length is set at 50 and a SciBERT-only model is used.

Metadata Choice & Feature Engineering: Both of these experiments relate to testing hypothesis **H2**. We first investigate the effects of choosing different metadata to enrich the citation context. For this work, we consider: Title only, Abstract only and Title and abstract together. We use SciBERT for both citation context and metadata encoders. We fix the metadata embedding as the second-to-last hidden layer of the SciBERT model.

We then fix the combination of metadata that shows most promise before trialling other methods of constructing metadata embeddings from the hidden layers of BERT models. We experiment with the last hidden layer, second-to-last hidden layer, the average of all 12 layers, and a concatenation of the last four hidden layers. Again, SciBERT is used for both citation context and metadata encoders. We carry out fine-tuning on the development dataset to avoid overfitting to the test data.

The best combination of feature construction and metadata is then trained on the full dataset for comparison to other models.

6 Results and Discussion

We refer to Table 3 for the most part in our discussion of results.

6.1 Reproducing Previous Work

BERT(-Only) and BERT-GCN results from (Chanwoo Jeong, 2019) are compare with our reproduced results. Whilst we achieved similar metrics for BERT-Only, we were not able to reproduce the marginal gains from the GCN, instead noting a slight reduction in performance (although perhaps a statistically insignificant difference). It is unclear whether their results are 'best' runs, or averaged over multiple repeats, which might account for the discrepancy if, for example, a fortunate initialisation during training of the GCN led to better results.

6.2 Choice of Citation Context Encoder

The application of SciBERT as citation context encoder leads to $\approx 14\%$ improvement in MAP/MRR and similar in recall metrics over baseline BERT and BERT-GCN, providing new state of the art on this dataset. The same cannot be said from using BioBERT as the context encoder, where slight differences above baseline performance are likely within the noise level of repeat runs. Two factors could explain SciBERT's performance over BioBERT. Firstly, SciBERT was pre-trained on roughly one fifth computer science papers whereas BioBERT contained no papers in this domain; the weights of pre-trained SciBERT is therefore naturally adapted to downstream tasks related to Computer Science and hence FullTextPeerRead. Second is that SciBERT has its own vocabulary tailored to scientific literature, only retaining 42% of the orig-

Table 4: Example Results: Contexts and top ranked predicted citations for BERT and SciBERT models with context sequences of 50 and frequency 5.

Ex.	Context	True Citation	Predicted BERT	Predicted SciBERT	Rank BERT	Rank SciBERT
1	del is trained end-to-end using the Adam optimizer [?] with a mini-batch size of 100.Fig. 1 shows four sa	Adam: A Method for Stochastic Optimization	Adam: A Method for Stochastic Optimization	Adam: A Method for Stochastic Optimization	1	1
2	or both better generalization and domain transfer. [?] proposed backtranslation as a way of using unlabel	Improving Neural Machine Translation Models with Monolingual Data	Frustratingly Easy Domain Adaptation	Improving Neural Machine Translation Models with Monolingual Data	394	1
3	c regression and Lasso optimization problems [?] took advantage of the sparsity inherent in models	Parallel Coordinate Descent for L1-Regularized Loss Minimization	Parallel Coordinate Descent for L1-Regularized Loss Minimization	Lasso Screening Rules via Dual Polytope Projection	1	324
4	uce the discrepancy between training and testing [?] .Given a pre-trained sequence generation model, an	Sequence-to-Sequence Learning as Beam-Search Optimization	Effective Approaches to Attention-based Neural Machine Translation	Sequence to Sequence Learning with Neural Networks	471	416

inal BERT vocabulary (Iz Beltagy, 2019). Keeping the general vocabulary of BERT in BioBERT thus produces suboptimal tokenization of scientific literature. Potentially a third reason is that BioBERT uses a cased vocabulary, generally shown empirically to have worse performance compared to uncased vocabulary (J. Devlin and Toutanova, 2018).

As with the baseline BERT-GCN model, we do not see any improvement over SciBERT-only with SciBERT-GCN.

Citation Recommendation Examples

Example citation recommendations in Table 4 can help us understand performance between BERT and SciBERT qualitatively. Recall there are $C = 489$ potential papers that can be recommended. Example 1 shows both models performing well on a straightforward recommendation, where the context includes key words 'Adam Optimizer' to match to the Adam optimization paper. In Example 2, SciBERT ranks the correct citation first whereas BERT ranks it much lower; the paper recommended by BERT still has relevance on the topic of domain adaptation, but less-complete tokenization of 'generalization' and 'backtranslation' may be the difference in recommending the correct paper. In Example 3, BERT now ranks the correct citation first, with SciBERT ranking it much lower; SciBERT here returns a paper with relevance regarding the topic of Lasso regularisation. In Example 4, both BERT and SciBERT rank the true citation poorly; again, SciBERT appears to recommend a paper of considerable relevance. This emphasizes potential problems when evaluating metrics such as MAP when only a single paper (the true cited paper) is regarded as relevant.

Citation Sequence Length

In both BERT-only and SciBERT-only models, additional context leads to improvement in performance across the range of sequence lengths explored (Figure 4). There is indication of diminishing returns, possibly because the relevance of context decreases with distance from the citation placeholder to the point that additional context just adds noise. There is no observed sensitivity difference between BERT and SciBERT models as the gap in performance is consistent with varying sequence length.

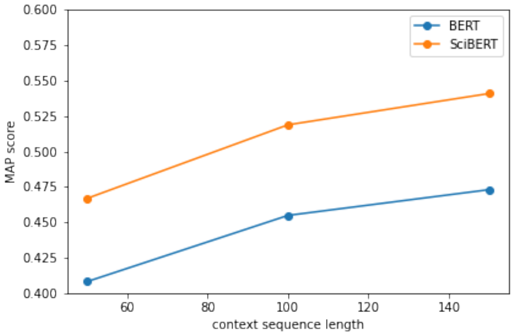


Figure 4: MAP score against citation sequence length for BERT and SciBERT models

6.3 Metadata Encoder

The metadata selected for encoding potentially shows an impact on performance in Table 5, albeit too small to be confident. The results might

Table 5: BERT-META MAP results for different feature engineering approaches and combinations of metadata. Results reported are on development dataset.

Feature Construction	Metadata	MAP
Baseline (N/A)	N/A	0.4019
Second-to-last Layer	Title + Abstract	0.3984
Second-to-last Layer	Abstract	0.3946
Second-to-last Layer	Title	**0.4006**
Last Layer	Title	0.4004
Average All Layers	Title	0.3956
Concat Last Four Layers	Title	0.3986

suggest that using the title alone to generate the embedding gives better performance than abstract only or a concatenation of title and abstract. We speculate that this could be due to limitations of BERT for long sequence embeddings (Reimers and Gurevych, 2019).

Likewise, experiments to construct the embeddings with different features from SciBERT layers did not provide an outright best method for improving the downstream task. The second-to-last or last layer may have the edge, but again these results are not yet tested for robustness. No method found appeared to beat a SciBERT-only model on the development dataset.

We choose our final BERT-META model to use only title embeddings constructed from the average across all tokens in the final layer. As suggested from the development performance, this model was unable to outperform the SciBERT-only approach. A limitation of BERT/SciBERT is that they do not naturally produce sentence/sequence embeddings, and the technique of averaging across all tokens may be inadequate for the citation recommendation task (Reimers and Gurevych, 2019; Pappagari et al., 2019). It may also be that other metadata such as author and venue are more informative than title or abstract and should be incorporated through other encoding methods (as they are unlikely to be contained within SciBERT's vocabulary).

6.4 Impact of Truncated Words

Modifying the citation context sequence to only contain whole words sees a small drop in performance against baseline SciBERT-only. Though unexpected, this may indicate that incomplete words contains information for the downstream task.

7 Conclusion

Our study uses the work of (Chanwoo Jeong, 2019) as a foundation for improving state of the art on the FullTextPeerRead dataset.

We have demonstrated that our proposal to encode citation context with a domain-specific BERT variant – SciBERT – achieves new state of the art on this dataset. However, investigations into enriching context embeddings with metadata embeddings, also constructed with SciBERT, have thus far been unable to further improve performance. We conclude that the feature-based approach used may be too general for the downstream task, and that a fine-tuning method to learn metadata embeddings may hold more promise for future work. The metadata encoder can be fine-tuned end-to-end simultaneously with the context encoder, either separately or as one in a Siamese-*style* [6] network (Bromley et al., 1993), the latter perhaps offering more rigidity to avoid severe overfitting to training data. Incorporating other available metadata such as venue and author, as suggested in (Dai, 2018) and (Ebesu and Fang, 2017), could prove more informative than title and abstract.

Our exploration of FullTextPeerRead emphasizes the challenge to find datasets that are of high quality, complete with metadata and full text, and of reasonable size. FullTextPeerRead still suffers from noise due to the difficulties of extracting full texts from PDFs, and is small when compared to recent datasets such as unarXiv (Saier and Farber, 2019). Qualitative analysis of model predictions in Table 4 exposes the issue of using evaluation metrics such as MAP when only considering one cited paper (the true cited paper) to be relevant; it does not consider scenarios where multiple papers are equally correct to cite. Being able to label multiple papers as relevant for each citation placeholder could help discriminate and understand performance differences better, but require expert annotations to judge relevance.

We find that citation context length is important to the downstream task, with longer contexts improving performance significantly. Such is its importance that discarding incomplete words results in a slight drop in performance. One might consider incorporating knowledge of placeholder position within a sentence to improve performance for a fixed context length. This adaptive context could include either more left context, or more right context depending on the placeholder's position.

[6] We loosely use the term 'Siamese' as we are considering two different inputs passed into the same network. However, unlike traditional Siamese networks, we would not aim to compare the vectors.

References

Caragea C. et al. 2014. Citeseerx: A scholarly big dataset. *Advances in Information Retrieval. ECIR 2014. Lecture Notes in Computer Science*, 8416.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a ”siamese” time delay neural network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, page 737–744, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Hyuna Shin Eunjeong Park Sungchul Choi Chanwoo Jeong, Sion Jang. 2019. A context-aware citation recommendation model with bert and graph convolutional networks. *Scientometrics*, arXiv:1903.06464v1.

Libin Yang; Yu Zheng; Xiaoyan Cai; Hang Dai; Dejun Mu; Lantian Guo; Tao Dai. 2018. A lstm based model for personalized context-aware citation recommendation. *IEEE Access*, 6.

Bhavana Dalvi Madeleine van Zuylen Sebastian Kohlmeier Eduard Hovy Roy Schwartz Dongyeop Kang, Waleed Ammar. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1.

Vahed Qazvinian Dragomir R. Radev, Pradeep Muthukrishnan and Amjad Abu-Jbara. 2013. The acl anthology network corpus. *Language Resources and Evaluation*, 47, pages 919–944.

T. Ebesu and Y. Fang. 2017. Neural citation network for context-aware citation recommendation. *Proc. 40th SIGIR Conf.*

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Qi He, Daniel Kifer, Jian Pei, Prasenjit Mitra, and C. Lee Giles. 2011. Citation recommendation without author supervision. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, page 755–764, New York, NY, USA. Association for Computing Machinery.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware citation recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 421–430, New York, NY, USA. Association for Computing Machinery.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Arman Cohan Iz Beltagy, Kyle Lo. 2019. Scibert: A pretrained language model for scientific text. *ACL Anthology*.

K. Lee J. Devlin, M.-W. Chang and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint*.

Jieh-Sheng Lee and Jieh Hsiang. 2019. Patentbert: Patent classification with fine-tuning a pre-trained BERT model. *CoRR*, abs/1906.02124.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained transformers for text ranking: BERT and beyond. *CoRR*, abs/2010.06467.

Adam Jatowt Michael Färber. 2020. Citation recommendation: approaches and datasets. *International Journal on Digital Libraries (2020) 21:375–405.*

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. *CoRR*, abs/1910.10781.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Tarek Saier and Michael Farber. 2019. Bibliometric-enhanced arxiv: A data set for paper-based and citation-based tasks.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. *CoRR*, abs/1904.01766.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.