Scientometrics (2020) 124:1907–1922
https://doi.org/10.1007/s11192-020-03561-y

A context-aware citation recommendation model with BERT and graph convolutional networks

Chanwoo Jeong[1] · **Sion Jang**[1] · **Eunjeong Park**[2] · **Sungchul Choi**[1]

Received: 6 September 2019 / Published online: 13 July 2020

Abstract
With the tremendous growth in the number of scientific papers being published, searching for references while writing a scientific paper is a time-consuming process. A technique that could add a reference citation at the appropriate place in a sentence will be beneficial. In this perspective, the context-aware citation recommendation has been researched for around two decades. Many researchers have utilized the text data called the context sentence, which surrounds the citation tag, and the metadata of the target paper to find the appropriate cited research. However, the lack of well-organized benchmarking datasets, and no model that can attain high performance has made the research difficult. In this paper, we propose a deep learning-based model and well-organized dataset for context-aware paper citation recommendation. Our model comprises a document encoder and a context encoder. For this, we use graph convolutional networks layer, and bidirectional encoder representations from transformers, a pre-trained model of textual data. By modifying the related PeerRead dataset, we propose a new dataset called FullTextPeerRead containing context sentences to cited references and paper metadata. To the best of our knowledge, this dataset is the first well-organized dataset for a context-aware paper recommendation. The results indicate that the proposed model with the proposed datasets can attain state-of-the-art performance and achieve a more than 28% improvement in mean average precision and recall@k.

Keywords Paper citation · Citation recommendation · BERT · Deep learning · Transformer · Graph convolution network

Mathematics Subject Classification 68U15

✉ Eunjeong Park
lucy.park@navercorp.com

✉ Sungchul Choi
sc82.choi@gachon.ac.kr

[1] TEAMLAB, Department of Industrial Engineering, Gachon University, Seongnam-si, Gyeonggi-do, Republic of Korea

[2] Papago, NAVER, Bundang-gu, Seongnam-si, Gyeonggi-do, Republic of Korea

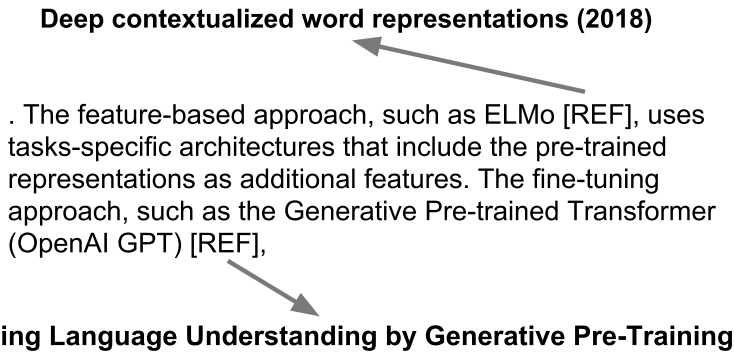


Fig. 1 An example manuscript with citation placeholders to find fittable references. This text is from BERT paper (Devlin et al. 2018)

Introduction

Have you ever had difficulty citing references while writing a scientific paper? Most artificial intelligence researchers may have thought about finding a solution for this problem a little easier at least once. As shown in Fig. 1, a possible solution is to automatically find the information that should be cited in a placeholder, such as [REF].

Finding a suitable scientific document—considering the surrounding text—that can be used as a placeholder is called "Context-aware citation recommendation" (He et al. 2010). The sentences on both sides of a placeholder are referred to as "context". A context-aware citation recommendation task is a type of supervised classification that is used to choose a paper suitable as the placeholder based on its contents. In addition to context, by considering the characteristics of a scientific document, the task has conducted with using author, title, citation, journal (or conference name), etc., which are metadata, or bibliometrics, of a scientific paper (He et al. 2011; Liu et al. 2015; Rokach et al. 2013; Tang et al. 2014; Bai et al. 2019; Moed 2010; Kim and Chen 2015). In recent years, there has been an increasing number of attempts to solve problems with the use of a deep neural net; this resulted in an increase in the number of articles (Huang et al. 2015; Ebesu and Fang 2017; Yang et al. 2018).

Even though the task has been a relatively constantly researched area, one of the most challenging aspects of this study is that there is no benchmarking dataset against which proper performance can be measured. In general, this task needs to use metadata along with the context surrounding the cited paper. To the best of our knowledge, suitable datasets have not been disclosed. Among the commonly used data, the ACL Anthology Network (AAN)[1] dataset does not provide paper sentences and metadata in a preprocessed form, and the DBLP dataset[2] only provides bibliographic information. In a recently published Huang et al. (2015), CiteseerX datasets[3] only provided context and citation information and did not provide meta information simultaneously. As a result, related studies have failed to use the same benchmarking dataset (Fig. 2).

[1] http://clair.eecs.umich.edu/aan/indexphp.
[2] https://dblp.uni-trier.de/xml/.
[3] https://psu.app.box.com/v/refseer.

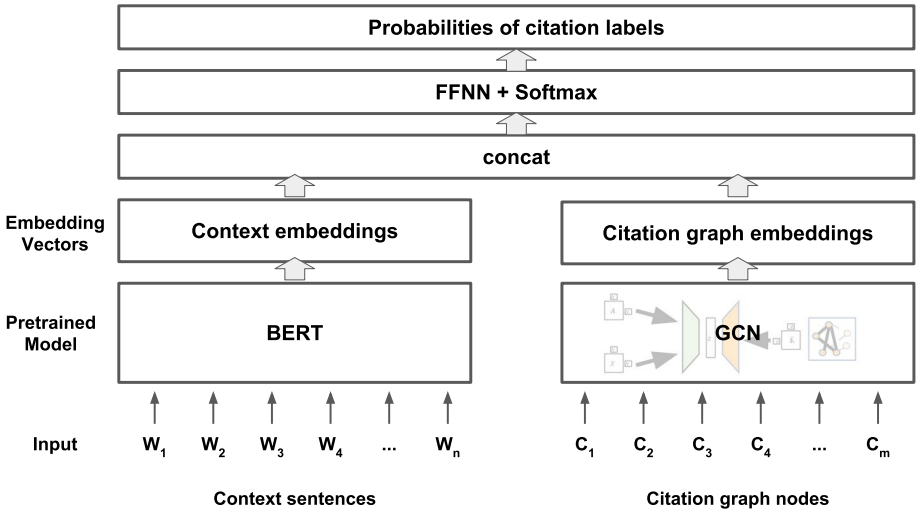


Fig. 2 A BERT–GCN model architecture

The purpose of this study is to provide datasets and state-of-the-art models suitable for the context-aware paper recommendation task research, and in turn provide researchers with an improved paper writing environment. The main contributions of this study are as follows: first, we built reproducible benchmarking datasets for the task. We preprocessed the existing AAN dataset (Radev et al. 2013, 2009; Dragomir et al. 2009) to fit the task, and constructed new dataset called FullTextPeerRead using PeerRead[4]Kang et al. (2018). Second, we constructed the state-of-the-art model for the task using BERT (Devlin et al. 2018) and graph convolution networks (GCN) (Kipf and Welling 2016). Because scientific papers contain textual contents data, and metadata that can be represented as a graph, we use BERT, which recently proved to have the highest performance level in the field of Natural Language Processing (NLP) for textual data, and GCN for network-based metadata. Finally, we investigated various factors to affect task performance through experiments.

Related work

This chapter describes the research history that the citation recommendation system and graph-based article citation relation embedding.

Context-aware recommendation

Unlike general citation recommendation based on similarity between documents, Context-aware citation is a study that finds highly relevant papers based on citation context or syntax. The following paragraph describes the flow related to the context-aware citation recommendation study. Context-aware Citation Recommendation (He et al. 2011) consists of

[4] https://github.com/allenai/PeerRead.

global citation and local citation context. Here, citation-recommendation was performed through a non-parametric probabilistic model. In addition, Neural Citation Network for Context-Aware Citation Recommendation (Ebesu and Fang 2017) was used through citation context and authors network convolution. In this paper, a citation recommendation list was studied by adding a cited paper title to a GRU decoder by applying an attention mechanism by combining citing and cited author network features in a citation context encoder. As a similar study, a study is conducted to recommend people based on celebrity quotes. QuoteRec: Toward Quore Recommendation for Writing (Tan et al. 2018) is a study that performs a task of Quote (citation) verses of celebrities. The person's name and Topic are encoded in the verse. In this paper, propose a model that proposes applying the author's information to a paper encoder through LSTM. From the point of view of the forward and backward positions of the citation context's quotation marks, A LSTM based model for personalized context-aware citation recommendation (Yang et al. 2018) separates the left 2 sentences and the right 3 sentences based on the quotation marks when encoding the citation context. This paper proposed a method to learn LSTM cells, merge LSTM cells, and learn with MLP. Also, in the paper encoder area, the author and venue information of the paper is applied to the paper encoder as in the previous study model.

Document relation encoder

Bibliography (author, title, abstract, venue, year) information that defines paper is a unique feature of paper. Bibliography information is also used in paper encoders in citation recommendations. Previously, the technique of encoding information of paper was mainly used by using information such as title and abstract in Doc2vec. This technique is effective in terms of individual paper, but it has limitations in containing the Citation organic relation. Therefore, recently, a technique using Graph reflecting the citation relationship was introduced. Learning convolutional neural networks for graphs (Niepert et al. 2016) introduced the conversion of graph information to grid information using the convolution technique. In addition, Semi-supervised classification with graph convolutional networks (Kipf and Welling 2016) proposed a method to classify all graph nodes through semi-supervised by utilizing the features of graph and graph nodes through graph convolution networks. Here, spectral graph convolution is applied to directly use CNN for graph information. Specifically, they represent GCN through layer-wise forward propagation. They also propose learning latent representation through unsupervised learning by grafting GCNs to Variational auto-encoders (VAE) through Variational Graph Auto-Encoders (Kipf and Welling 2016).

Proposed dataset

Dataset overview

We constructed two new datasets for the context-aware citation recommendation task. We suggested revisions of the existing datasets, AAN (Radev et al. 2013) and FullTextPeerRead which is an expansion of PeerRead dataset (Kang et al. 2018). AAN and PeerRead datasets have well-organized bibliometric information. The PeerRead dataset mainly provides peer reviews of submitted papers in top-tier venues of the artificial intelligence field, along with bibliometric information. Since all existing datasets of the task lack information

Table 1 Dataset description

Dataset name	AAN	FullTextPeerRead
# of total papers	7073	4898
# of base papers	5576	3761
# of cited papers	2417	2478
# of citation context	12,125	17,247
# Paper published year	1965–2015	2007–2017

Because the AAN policy prohibits disclosure of modified data, we cannot disclose it. We will open the dataset after receiving a grant from the AAN

http://bit.ly/2IDam8J

http://bit.ly/2Srkdht

in the citation context, our focus is on gathering context information with metadata. Therefore, the AAN and PeerRead datasets were reprocessed to create the dataset.

Data acquistion

We used arXiv Vanity[5] to create the dataset. arXiv Vanity is a site that converts a latex-based PDF file to HTML document. Our goal is to extract the context information on both sides of the citation symbol, such as Devlin et al. (2018) or [Choi et al. 2016], together with the reference paper information. To do this, we parsed latex into HTML via arXiv Vanity, and used a regular expression to recognize the citation symbol in the document. Next, the sentences on both sides of the citation symbol were stored on a database with reference article information. We stored the collected information together with the existing metadata and build it into a database.

The data actually collected was noisy because the latex documents were not consistent format. After we automatically collected the necessary data, we removed the noisy data manually. For example, in the case of CiteseerX[6], the citation symbol corresponding to the placeholder is left in the context and data is provided. However, in this case, the placeholder text itself was used for overfitting learning, so the text can be used to tell the correct answer. Exactly, the reminded placeholder can be used as crucial evidence of prediction. Therefore, our final work was post-processed manually because this noise may remain when mechanically collected.

Statics of datasets

Finally, the statics of the dataset we built are as shown in Table 1. The number of extracted datasets were less than the original number of AAN or PeerRead dataset because we needed to remove paper the PDFs that did not use latex or were very noisy from being processed with arXiv Vanity. In Table 1 below, Total "# of base papers" is a paper that cites other researches. We have metadata information of the paper that is used as an input for the classification task. "# of cited papers" is a cited paper. In addition, we extracted

[5] https://www.arxiv-vanity.com/.

[6] https://psu.app.box.com/v/refseer.

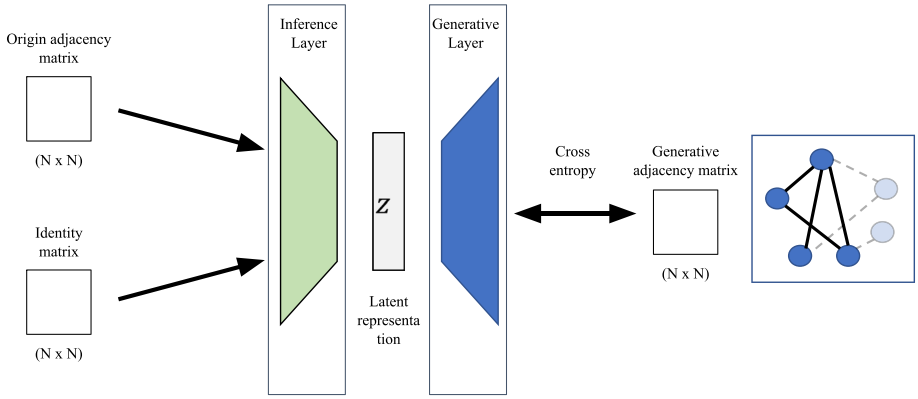


Fig. 3 A architecture of Variational Graph AutoEncoder

paragraph units on both sides of the citation symbol, and "# of citation context" means the sum of the number of sentences which are in the extracted paragraph. Further, "# of total papers" refers to the total number of papers covered by base paper and cited paper, excluding duplicates.

A BERT–GCN model for context-aware citation recommendation

Model overview

We construct the context-aware citation recommendation model using BERT (Devlin et al. 2018) and GCN (Kipf and Welling 2016). BERT is one of the highest performing pre-trained models for NLP learning representation. We expect that the learning presentation of context sentences, through pre-trained BERT, will achieve a high performance. Scientific data, such as papers, also contain various metadata, in addition to textual data. We use the GCN model to represent the citation relationship between papers and to extract a learning representation of them. We thought that the information extracted based on these two models would be useful for the new citation recommendation. So far, there are not many papers that applied bert model to citation and graph information or gcn to text information. Therefore, bert based text and gcn based grah are considered to be the most suitable models.

As shown in Fig. 3, we construct a context encoder to extract textual embedding, using BERT, and a citation encoder to extract graph embedding from GCN. Each encoder is pre-trained with context data, and citation graph data is extracted from the paper. Then data is inserted into the pre-trained models, and concatenated embeddings are calculated by each encoder. Finally, after passing the concatenated vectors to a feedforward neural net, the softmax output layer is generated, and cross entropy is adopted as a loss function for training. As a result, the model predicts the probability for citation. The more the citation probability converges to 1, the higher the citation relationship between the input sentence and the paper appears. In addition, by sorting the results of the citation probability of the paper in descending order, the citation ranking is arranged in high order.

The structure of the proposed model is linked to the baseline CACR (Yang et al. 2018). CACR has both a paper encoder and a citation context encoder. CACR demonstrates the performance of the State-of-The-Art (SOTA) as the most recent context-aware citation recommendation model using an AAN dataset and an LSTM model. In the CACR model, a paper encoder was constructed using author, venue, and abstract information in the paper. Our model constructed the citation encoder with GCN solely using citation information.

Citation encoder

The citation encoder conducts unsupervised learning for citation, linking a prediction with the GCN-based Variational Graph Auto-Encoders (VGAE) model (Kipf and Welling 2016) by using citation relationships between papers as input values. When paper information is used as input to a pre-trained GCN, the model returns the relational learning representation as the embedding vector. VGAE can capture the latent learning representation of graph data.

In existing research, it has been a challenge to convey the citation relationship of a paper as Doc2Vec (Le and Mikolov 2014) has been used to encode paper information and summarize it after embedding the learning of the individual meta-information. Our citation encoder complemented this by using citation linking prediction information as a citation prediction feature.

Graph convolutional network (GCN) layer

In our model, the role of the GCN layer is to abstract the citation network graph information through a convolutional network. The GCN layer is used as an inference model of VGAE. The formula of the GCN layer for VGAE is shown in Eq. 1.

$$GCN(X, A) = \tilde{A}ReLU(\tilde{A}XW_0)W_1 \tag{1}$$

The proposed model consists of two GCN layers. The layer uses two matrices as input: the identity matrix X and the adjacency matrix A, which is an N by N matrix. N is the number of input paper. Learned from the first GCN layer, layer parameter W^0 is used as the weight matrix for the second layer. Each layer uses layer-wise propagation.

$\tilde{A}$ is a normalized adjacency matrix based on the diagonal degree matrix D, as shown in Eq. 2.

$$\tilde{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \tag{2}$$

Variational Graph AutoEncoder (VGAE)

VGAE is a model that applies the unsupervised learning method of the Variational Autoencoder (Rezende et al. 2014) to the graph structure using GCN. VGAE learns the latent representation by minimizing the cost between inference model and the generative model as shown in Eq. 3.

$$L = \mathbb{E}_{q(Z|X,A)}[\log p(A|Z))] - KL[q(Z|X, A)||p(Z))] \tag{3}$$

The inference layer of VGAE learns the representation Z by decreasing the KL-divergence loss between the normal distribution from the GCN layer result and the Gaussian normal distribution, as shown in Eq. 4.

$$\begin{aligned} q(Z|X,A) &= \prod {}_{i=1}^{N} q(z_i|X,A), q(z_i|X,A) \\ &= \mathcal{N}(z_i|\mu_i, diag(\sigma_i^2))) \end{aligned} \tag{4}$$

As a next step, the generative layer learns an adjacency matrix based on the representation matrix Z of the interference layer. The latent variable z_i, z_j is the inner product value of document i and j. An adjacency matrix is generated based on the latent variable through the inner product between paper vectors, as shown in Eq. 5. The generative model defines the representation matrix Z by reducing the difference between the adjacency matrix A generated by the generative model and the actual adjacency matrix.

$$\begin{aligned} p(A|Z) &= \prod {}_{i=1}^{N} \prod {}_{j=1}^{N} p(A_{ij}|z_i, z_j), p(A_{ij} = 1|z_i, z_j) \\ &= \sigma(z_i \top z_j) \end{aligned} \tag{5}$$

Context encoder

BERT model suggests 4 usage modes based on the task. Related contents include CLS–SINGLE, CLS–PAIR, Name Entity Recognition, and Question Answering. We applied BERT's CLS–PAIR model in our proposed model within context encoder. For the Title and Abstract Feature used in the our proposed model, it was possible to represent a useful paper representation by applying CLS–PAIR. A fully connect layer was built on the transformer output of the pre-trained model and used as a representation through this. As a fine-tuning model input value, The sentence on the left (the citation placeholder left) and the right (the citation placeholder right) are put as input sentence pairs. Also, BERT is divided into Base and Large models according to the size of the model. We use base BERT model (English Pre-train, 12-layer, 768-hidden, 12-heads, 110M parameters)[7] for the proposed model. Finally, the context embedding is used as Representation by stacking the Fully Connect Layer in the hidden state above the Transformer output of the [CLS] token, such as Classification Fine-tuning.

Experiments

Experiments overview

We compare the proposed model with CACR (Yang et al. 2018), one of the existing SOTA models, with a focus on performance. We use the AAN and FullTextPeerRead (FTPR) datasets in our experiments and use Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), and Recall@K as evaluation metrics. The purpose of our experiments is to investigate the following topics, including the overall performance of our model.

[7] https://github.com/google-research/bert.

- We compare the performance of the proposed model with the existing SOTA, the CACR model, to gauge the outperformance of BERT and GCN over conventional model.
- We investigate differences in performance between models using BERT and GCN. We used BERT for textual data and GCN for graph data. We analyze the impact of each models on overall performance.
- When using this model in a practical environment to recommend papers, we want to verify that the model can recommend papers solely by looking at the sentence on the left side of a citation symbol. In general, if researchers are writing a paper, researchers expect to recommend reference papers based on the contents which have been written so far without the entire document. We want to check whether our model is available in this situation.
- We check the performance of the model according to the length of textual data. When using BERT, we check whether sentences, that are distant from the citation symbol, are used as noise or useful information
- We measure performance according to paper occurrence in the aggregate dataset. Citation of specific papers is rare; however, we need to understand how our model performs when it happens.

Experiments setting

Experimental dataset

In the experiment, the AAN dataset used data published before 2014, whereas the FullTextPeerRead dataset comprised paper data published before 2018. After interpreting the data, with sufficient context, and the metadata among datasets, AAN and FullTextPeerRead yielded 6500 and 4898 papers, respectively. The datasets were divided into two parts: the AAN dataset used 5806 pre-2013 papers for the training set, and the remaining 973 others for the test set prior to 2013, and the test set 973 for the test set. Regarding the FullTextPeerRead dataset, 3411 pre-2017 papers were used for the training set, and 2559 papers from 2017 were used for the test set. We applied two text lengths, namely 50 and 100, to the citation context features, and measured the related shift in performance. Furthermore, in order to test performance according to the text input characteristics, we conducted an experiment comparing single and pair context characteristics. For this purpose, we used a single context consisting of a 100 text length on the left side of a citation placeholder, and the pair context consisting of a 100 text length on each side of the citation placeholder.

Evaluation metrics

For experimental evaluation, we use MAP, MRR, and Recall Top@K, which are general metrics for information retrieval. The MAP measures average precision reflecting the rank position regarding the retrieval list. This indicator is based on the position of the corresponding label values for the K recommendation list, and we measured the indicator with $K = 30$. The MRR indicator is defined as identifying the location of the first occurrence of the actual labels in the recommendation list. Finally, Recall Top@K is defined as an indicator of the actual label hit ratio in a Top@K recommendation list. For this, we evaluate the Recall index through $K = 5, 10, 30, 50, 80, 100$.

Parameter setting

We extracted the embedded context vectors and document vectors from the BERT and GCN layers built in a separate learning process. In BERT, the number of multi-head attention is 12, the encoder stack is 12, the total number of epochs for learning is 30, the batch size is 16, and the Adam optimizer is used. The learning rate is 2e−5, epsilon is 1e−6, beta#1 is 0.9, beta#2 is 0.999, and the weight decay rate is 0.01. We set the sequence length maximum to 128, padding 0 if the length is shorter than 128, and the hidden size is 768.

With regard to GCN, the number of epochs is 200, the first hidden dimension is same as the document size and the second hidden dimension is 768, the batch size is same as the total document size (full-batch gradient descent), the optimizer is the Adam optimizer Kingma and Ba (2014), and the learning rate is 0.01. And regard to Doc2vec, the number of epochs is 50, hidden size is 300, text window is 10, min count is 5 and alpha is 0.025

Experiments results

Baseline comparison

As shown in Table 2, our model delivers a significant performance improvement over the existing CACR. Compared with the SOTA model, all our models showed a three times performance improvement in MAP, MRR, and Recall@K indices, approximately. In particular, Recall@5, which only sees five retrieval citation papers, is a significant improvement.

In this experiment, both our model and CACR were used solely for papers with a minimum fequency of five citations, and the learning was conducted by considering fifty words on both sides of the citation symbol (Table 3).

We compared the performance by independently reproducing the Python code related in the CACR paper. There was no detailed experimental information such as frequency in the actual paper. Since no frequency was mentioned, we assume that the performance described in the CACR paper is based on a frequency of one, and we compare our performance with the performance described in the paper of CACR as shown in Table 2. For MAP, MRR, and Recal@10, our model outperform, but after Recall@10, it does underperform. Based on experience, we guessed this to be a phenomenon that occurs when the classification label value is returned with a high frequency of cited papers.

Impact of BERT and GCN

In all cases, when the GCN was added, the performance of the model improved. Of interest is the impact that GCN has on the performance of BERT, and BERT-left alone. As shown in Fig. 4, the difference in recall performance between BERT + GCN-left and BERT-left differs from the difference in performance difference between BERT + GCN and BERT. BERT-left has half the input of context sentence of BERT. Hence, the impact of GCN is greater when the value of the input words is small.

Impact of the left side context alone

From the viewpoint of an actual researcher writing a paper, we check the performance when using the context sentence on the left side of the citation symbol alone. As shown in Table 2, the model using the left context alone has a performance that is approximately

Table 2 MAP, MRR, and Recall@K scores for a frequency of over five citations and fifty pair context sentences. Bold means that performance is highest in the metric in a comparison between models

Dataset	Model	MAP	MRR	Recall@5	Recall@10	Recall@30	Recall@50	Recall@80
AAN	BERT–GCN	**0.6189**	**0.6036**	**0.6736**	**0.7109**	**0.7814**	**0.8162**	**0.8538**
	BERT–GCN-left	0.5967	0.5818	0.6459	0.6843	0.7506	0.785	0.8245
	BERT	0.6118	0.5971	0.6593	0.6976	0.7645	0.81	0.8257
	BERT-left	0.5928	0.5789	0.6364	0.6678	0.7379	0.7793	0.8203
	CACR (Yang et al. 2018)	0.2893	0.2917	0.3861	0.4531	0.5799	0.6573	0.721
	Doc2vec	0.0157	0.0170	0.0195	0.0313	0.0619	0.0619	0.1133
FullTextPeerRead	BERT-GCN	**0.4181**	**0.4179**	**0.4864**	**0.5291**	**0.6036**	**0.6495**	**0.6994**
	BERT–GCN-left	0.3883	0.388	0.4455	0.4815	0.5539	0.5991	0.6499
	BERT	0.4152	0.415	0.4801	0.52	0.5926	0.6366	0.6887
	BERT-Left	0.3823	0.3821	0.4391	0.4755	0.5459	0.5885	0.6392
	CACR (Yang et al. 2018)	0.1551	0.1549	0.2154	0.2761	0.4128	0.4794	0.5516
	Doc2vec	0.0577	0.0578	0.0857	0.1200	0.2037	0.2037	0.3123

Table 3 MAP, MRR, Recall@K scores with a frequency of over one citation, 200 pair context sentences and CACR (Yang et al. 2018) metric result. Bold means that performance is highest in the metric in a comparison between models

Dataset	Model	MAP	MRR	Recall@5	Recall@10	Recall@30	Recall@50	Recall@80
AAN	BERT–GCN	**0.4516**	**0.4323**	**0.4948**	**0.532**	0.5839	0.6118	0.6335
	CACR (Yang et al. 2018)	0.301	0.335	–	0.474	**0.615**	**0.653**	**0.712**

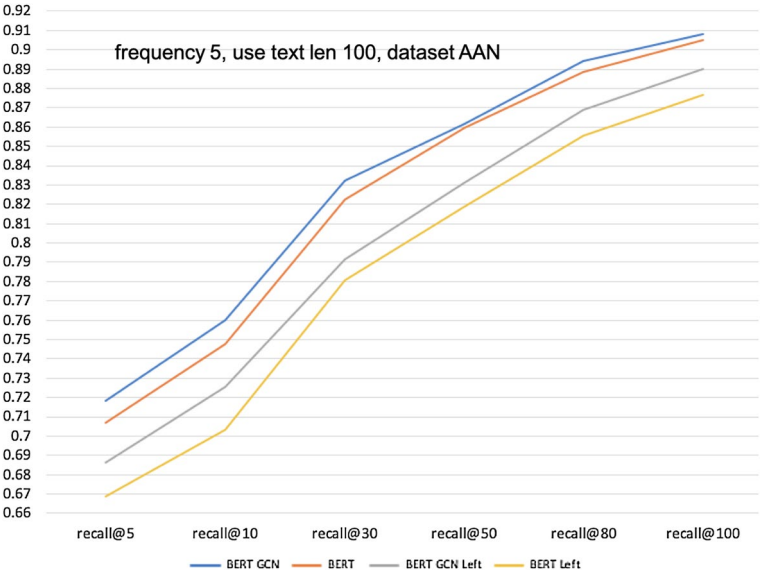
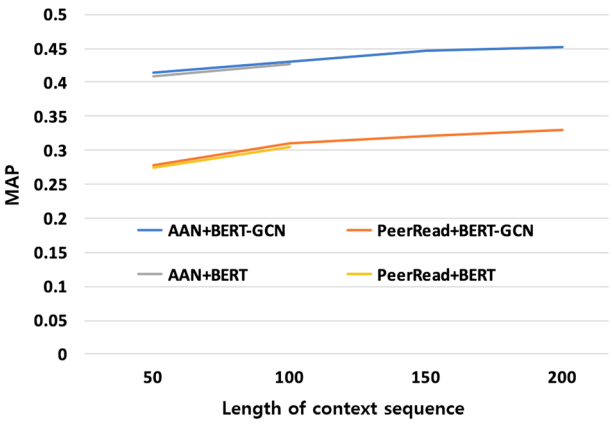


Fig. 4 The effects of BERT, GCN and BERT-left

Fig. 5 The shift in performance according to the length of sentence context when frequency is one



0.03 lower than the model using the entire context for the context encoder. In other words, it is helpful to consider the entire context on both sides of the quotation mark. However, adding GCN to the BERT model considering the left context alone delivers performance improvements over adding GCN to the BERT model that considers both contexts, as the network information of the GCN can be helpful when textual information is not sufficient.

Effect of length of the context sequence

We have already confirmed a shift in performance tied to the length of the input context sentences. However, how much more can performance improve as the number of sentences increases? As shown in Fig. 5, when the context length becomes a hundred or more, model performance becomes less impacted by comparative context length. The context sentence length is relevant, but beyond a certain point, its impact is actually reduced.

Effect of paper citation frequency

Finally, we examine the impact of citation frequency on performance. The results of the experiment with citation frequencies one, three, and five show that performance improves when citation frequency is higher, as shown in Table 5. In general, papers not cited are not used for learning and can be processed as sparse data even at the time of testing. We believe that the learning data should be refined according to citation frequency for a well-functioning, high-performance model.

Recommendation examples

Actual recommendation examples using our model can be found in Table 4. In these examples, we observe that one of the biggest effects of GCN is that it is possible to get a higher relevance similarity of cited paper through GCN for similar articles when we look at textual data alone. Expanding the use of GCN, would not only target similarities in the content of the papers but also citation information in the previous thesis or citation information in the current thesis. With regard to the use of GCN, the representation learning information concerning citation is generated by identifying not only a similarity in the content of the citing paper but also through the citation information of the previous paper by the author or the information in the cited paper of the current article. We think this information is very helpful to improve the performance of the context-aware citation recommendation process (Table 5).

Conclusion

The proposed model for context-aware citation recommendation task delivers a significant improvement in MAP, MRR and Recall@K over the existing model. The basis for the breakthrough performance improvement is that the BERT model, which has performed well in recent NLP tasks, is adapted to our context-aware framework. Through the context encoding via BERT, our framework improves the representation learning of the context side. In addition, we apply VGAE, which comprises a GCN layer according to graph data to mitigate over-fitting to local contexts when BERT is applied alone. The VGAE applied to our framework citation encoder processes the paper citation network graph data into

Table 4 A comparison of ground truth with the top five recommended citation lists

Citation context	Approaches	Top-3 system
... inbold means significant ly better than the baseline according to [?] or p value less than 0.01.baseline SMT system. The decoding ...	BERT–GCN	1. Statistical Significance Tests For Machine Translation Evaluation (O)
		2. Bleu: A Method For Automatic Evaluation Of Machine Translation (X)
		3. On Coreference Resolution Performance Metrics (X)
	BERT	1. Minimum Error Rate Training In Statistical Machine Translation (X)
		2. Statistical Phrase-Based Translation (X)
		3. Statistical Significance Tests For Machine Translation Evaluation (O)
	CACR	1. TnT—A Statistical Part-Of-Speech Tagger (X)
		2. Stochastic Inversion Transduction Grammars And Bilingual Parsing Of Parallel Corpora (X)
		3. A Maximum Entropy Model For Part-Of-Speech Tagging (X)
Many researchers have atte mpted to make use of cue p hrases especially for segme ntation both in prose [?]	BERT–GCN-left	1. TextTiling: Segmenting Text Into Multi-Paragraph Subtopic Passages (O)
		2. Multi-Paragraph Segmentation Of Expository Text (X)
		3. Discourse Segmentation Of Multi-Party Conversation (X)
	BERT-Left	1. Multi-Paragraph Segmentation Of Expository Text (X)
		2. Advances In Domain Independent Linear Text Segmentation (X)
		3. TextTiling: Segmenting Text Into Multi-Paragraph Subtopic Passages (O)

Table 5 A comparison of shifts in performance based on cited paper frequency

Model	Frequency 1	Frequency 3	Frequency 5
BERT GCN	0.4467	0.6063	0.6736
BERT (Devlin et al. 2018)	0.4423	0.5971	0.6593

a paper latent representation. The combination of the encoded paper network and the encoded context is regularized, resulting in a performance increase over a BERT-based model.

Regarding the context-aware citation recommendation study, existing datasets are not up-to-data and there is no clear context detection. To address this problem, we devised and released the FullTextPeerRead dataset. The proposed dataset comprises updated papers to an extent including papers up-to 2017, and provides a method to readily and accurately extract context metadata and has a well-organized perspective.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) Grant and funded by the Korean Government (No. NRF-2015R1C1A1A01056185 and NRF-2018R1D1A1B07045825).

References

Bai, X., Zhang, F., & Lee, I. (2019). Predicting the citations of scholarly paper. *Journal of Informetrics*, *13*(1), 407. https://doi.org/10.1016/j.joi.2019.01.010.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv e-prints

Dragomir, B. G. P. M., Radev, R., & Thomas, J. M. (2009). A bibliometric and network analysis of the field of computational linguistics. *Journal of the American Society for Information Science and Technology*.

Ebesu, T., Fang, Y. (2017). In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, SIGIR '17 (pp. 1093–1096). New York, NY: ACM. https://doi.org/10.1145/3077136.3080730.

He, Q., Kifer, D., Pei, J., Mitra, P., & Giles, C. L. (2011). In *Proceedings of the fourth ACM international conference on web search and data mining* (pp. 755–764). ACM

He, Q., Pei, J., Kifer, D., Mitra, P., & Giles, L. (2010). In *Proceedings of the 19th international conference on World wide web* (pp. 421–430). ACM

Huang, W., Wu, Z., Liang, C., Mitra, P., & Giles, C. L. (2015). In *Proceedings of the twenty-ninth AAAI conference on artificial intelligence* (pp. 2404–2410). AAAI Press.

Kang, D., Ammar, W., Dalvi, B., van Zuylen, M., Kohlmeier, S., Hovy, E., & Schwartz, R. (2018). In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: Human language technologies, Volume 1 (Long Papers)* (Vol. 1, pp. 1647–1661).

Kim, M. C., & Chen, C. (2015). A scientometric review of emerging trends and new developments in recommendation systems. *Scientometrics*, *104*(1), 239. https://doi.org/10.1007/s11192-015-1595-5.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv e-prints

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907

Kipf, T. N., & Welling, M. (2016). Variational graph auto-encoders, NIPS Workshop on Bayesian Deep Learning.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. arXiv preprint arXiv:1405.4053

Liu, H., Kong, X., Bai, X., Wang, W., Bekele, T. M., & Xia, F. (2015). Context-based collaborative filtering for citation recommendation. *IEEE Access*, *3*, 1695. https://doi.org/10.1109/ACCESS.2015.2481320.

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, *4*(3), 265. https://doi.org/10.1016/j.joi.2010.01.002.

Niepert, M., Ahmed, M., & Kutzkov, K. (2016). Learning convolutional neural networks for graphs. arXiv:1605.05273

Radev, D. R., Muthukrishnan, P., & Qazvinian, V. (2009). In *Proceedings, ACL workshop on natural language processing and information retrieval for digital libraries*. Singapore

Radev, D., Muthukrishnan, P., Qazvinian, V., & Abu-Jbara, A. (2013). The ACL anthology network corpus, language resources and evaluation pp. 1–26. https://doi.org/10.1007/s10579-012-9211-2.

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). In *Proceedings of the 31st international conference on international conference on machine learning—Volume 32* (JMLR.org), ICML'14 (pp. II–1278–II–1286). http://dl.acm.org/citation.cfm?id=3044805.3045035

Rokach, L., Mitra, P., Kataria, S., Huang, W., & Giles, L. (2013). A supervised learning method for context-aware citation recommendation in a large corpus. INVITED SPEAKER: Analyzing the Performance of Top-K Retrieval Algorithms, 1978.

Tang, X., Wan, X., & Zhang, X. (2014). In *Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval*, SIGIR '14 (pp. 817–826). New York, NY: ACM. https://doi.org/10.1145/2600428.2609564.

Tan, J., Wan, X., Liu, H., & Xiao, J. (2018). Quoterec: Toward quote recommendation for writing. *ACM Transactions on Information Systems*, *36*(34), 1.

Yang, L., Zheng, Y., Cai, X., Dai, H., Mu, D., Guo, L., et al. (2018). A LSTM based model for personalized context-aware citation recommendation. *IEEE Access*, *6*, 59618. https://doi.org/10.1109/ACCESS.2018.2872730.