Neural Citation Network for Context-Aware Citation Recommendation

Travis Ebesu
Department of Computer Engineering
Santa Clara University
Santa Clara, CA 95053, USA
tebesu@scu.edu

Yi Fang
Department of Computer Engineering
Santa Clara University
Santa Clara, CA 95053, USA
yfang@scu.edu

ABSTRACT

The accelerating rate of scientific publications makes it difficult to find relevant citations or related work. Context-aware citation recommendation aims to solve this problem by providing a curated list of high-quality candidates given a short passage of text. Existing literature adopts bag-of-word representations leading to the loss of valuable semantics and lacks the ability to integrate metadata or generalize to unseen manuscripts in the training set. We propose a flexible encoder-decoder architecture called Neural Citation Network (NCN), embodying a robust representation of the citation context with a max time delay neural network, further augmented with an attention mechanism and author networks. The recurrent neural network decoder consults this representation when determining the optimal paper to recommend based solely on its title. Quantitative results on the large-scale CiteSeer dataset reveal NCN cultivates a significant improvement over competitive baselines. Qualitative evidence highlights the effectiveness of the proposed end-to-end neural network revealing a promising research direction for citation recommendation.

CCS CONCEPTS

•Information systems →Information retrieval; •Computing methodologies →Neural networks;

KEYWORDS

Citation Recommendation, Deep Learning, Neural Machine Translation

1 INTRODUCTION

Authors establish credibility, honesty, and authority by providing accurate and relevant citations. The vast plethora of scientific literature makes searching for relevant work time consuming and highly keyword dependent. On the other hand, following the proceedings of well-known conferences restricts the scope of related work. Ideally, we desire a personalized, curated list of high-quality recommendations. We focus on the task of context-aware citation recommendation, where given a citation context (query), we recommend a list of high-quality candidate papers to fill the citation placeholder. A citation context comprises a small window of words surrounding a placeholder denoting where the citation should appear [2, 6–9]. We assume the surrounding text of a placeholder provides a short and concise summary of the paper's content.

Traditional information retrieval techniques rely heavily on keyword overlap, but identifying the critical structures in abstract ideas requires additional levels of semantic relations. For example, "deep learning" was previously known as "cybernetics" in its infancy and "connectionism" in its second resurgence [5]. As language evolves over time, new terms emerge while others become less frequently used. Similarly, the denotative meaning of words are generally fixed, perhaps more importantly, the connotative meaning changes throughout time. The words "deep" and "learning" treated independently as a bag-of-words lacks conceptual interpretation but modeling the conditional probability of the words together produces a clear concept. The word usage between the content in the citation context and corresponding cited document lead to a vocabulary gap [7–9] causing a mismatch between keywords leading to poor performance with standard information retrieval (IR) methods. In addition, existing methods cannot easily incorporate metadata without additional feature engineering or explicitly linked data [2].

We propose Neural Citation Network (NCN)[1] an encoder-decoder framework inspired by the success of neural machine translation (NMT) [1, 3, 10] which can learn relations between parallel pairs of variable-length text. Consequently, NCN is capable of characterizing the semantic composition of citation contexts and corresponding cited documents title by exploiting author relations. The encoder capitalizes on the computational advantages of a max time delay neural network [4] while the decoder leverages the capacity of recurrent neural networks (RNN) influenced by both the author networks and attention mechanism. As each composer of literature has her own writing style, grammatical structure, word usage and citation preference. NCN leverages these associated attributes with each author by utilizing only their name, producing significant performance gains. Furthermore, NCN can generalize to new papers not present in the training set. To the best of our knowledge, no prior work has addressed citation recommendation with the encoder-decoder framework. Experimental results on the CiteSeer dataset demonstrate NCN produces a significant improvement Recall, Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain (NDCG) over baseline methods. Qualitative results demonstrate the effectiveness of the proposed end-to-end neural network.

SIGIR '17, Shinjuku, Tokyo, Japan
© 2017 Copyright held by the owner/author(s). Publication rights licensed to ACM. 978-1-4503-5022-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3077136.3080730

[1] Source code: https://github.com/tebesu/NeuralCitationNetwork.

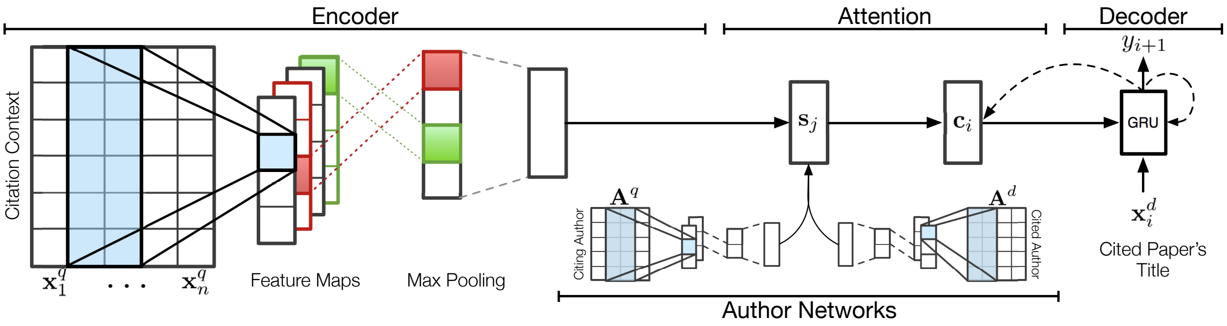

Figure 1: The proposed architecture of Neural Citation Network (NCN) with the attention mechanism and author networks. The dashed arrows represent recurrent dependencies.

2 RELATED WORK

Citation recommendation spans a variety of methodologies such as traditional IR, topic modeling, Restricted Boltzmann Machines, collaborative filtering, statistical machine translation (SMT) and neural networks [2, 8]. Due to space limitations, we focus on the latter two being the most relevant to our work. In SMT, a translation model treats the citation context and cited document content as parallel sequences [6, 7, 9]. The objective is to learn an alignment or transition probability the given citation context requires a citation. Lu et al. [9] learn an alignment from the citation context and the corresponding document's text, demonstrating improved performance over information retrieval methods when aligning to the shorter abstract rather than the full body of text. Similarly, Citation Translation Model (CTM) [7] treats each cited document as a token aligning the citation contexts to this reference. In order to address the noisy alignment between citation contexts and documents, He et al. [6] leverage topical information in their SMT model. More recently, Huang et al.[8] learned a distributed word representation of the citation context and the associated document embedding via a feedforward neural network. A comprehensive survey on citation recommendation can be found in [2].

NMT provides a general framework to address parallel pairs of arbitrary length sequences, where the source sequence is encoded to a fixed length representation followed by a decoder translating this representation to the target sequence conditioned on all previous states one token at a time. The encoder and decoder functions are application specific, in machine translation RNNs are typically used for both the encoder and decoder [1, 3] while in imaging captioning the encoder may be represented as a Convolutional neural networks (CNN) [10]. Bahdanau et al. [1] propose adding an alignment mechanism or attention model to the encoder-decoder framework alleviating the bottleneck placed on the encoder function to represent the entire source sequence.

CNNs demonstrate competitive performance to RNNs on natural language processing (NLP) tasks yet computationally cheaper by exploiting parallelism. In particular, the max time delay neural network (TDNN) [4] architecture performs a 1-dimensional convolution over a window of words constructing feature detectors followed by a max-pooling layer to extract relevant features from each sequence (time) simultaneously producing a fixed length representation.

3 NEURAL CITATION NETWORK

The proposed model is based on the encoder-decoder architecture with the attention mechanism [1] to integrate complementary author information and learn rich feature representations.

3.1 Encoder

In our encoder we leverage the TDNN [4] a CNN variant designed to capture long-term dependencies with a 1-dimension convolution over all possible word windows for a given context. A nonlinear projection coupled with max-pooling extracts rich feature representations from each convolved word window. Specifically, given a citation context of length n, let $\mathbf{x}_t^q$ be a g dimensional word embedding corresponding to the t^{th} word in the citation context and $\mathbf{x}_{1:n}^q = \mathbf{x}_1^q \oplus \ldots \oplus \mathbf{x}_n^q$ denote the concatenation of the embeddings from 1 to n. A convolutional filter $\mathbf{w} \in \mathbb{R}^{l \cdot g}$ slides over l words or regions at a time over all possible window lengths $\{\mathbf{x}_{1:l}^q, \mathbf{x}_{2:l+1}^q, \ldots, \mathbf{x}_{n-l+1:n}^q\}$, see Figure 1. We define the convolutional layer as:

$$o_k = \text{ReLU}(\mathbf{w}^\top \mathbf{x}_{k:k+l-1}^q + b_k); \qquad \hat{o} = \max\{o_1, \ldots, o_{n-l+1}\}$$

where ReLU is the nonlinear activation function $\max(0, x)$ and o_k is the k^{th} feature map, $\mathbf{o} \in \mathbb{R}^{n-l+1}$. The max-pooling over time yields a scalar representing the relevant feature $\hat{o}$ detected for the given set of feature maps subsequently converting the variable length sequence to a fixed one. In order to capture more complex relations the process is repeated p times with different filter weights yielding $\hat{\mathbf{o}}_j \in \mathbb{R}^p$. Finally, a fully connected layer allow interactions between the various phrase level feature maps extracted from the max-pooling layer, leading to:

$$\mathbf{s}_j = \tanh(\mathbf{U}_{s_j} \hat{\mathbf{o}}_j + \mathbf{b}_{s_j}) \tag{1}$$

where the TDNN aims to project the raw citation context $\mathbf{X}^q$, to a fixed summary representation $\mathbf{s}_j$ over feature maps of the j^{th} sliding region size of l_j. The final transformation $f(\mathbf{X}^q)$ applies a set of variable region size filters $L = \{l_1, \ldots, l_{|L|}\}$ to capture different granularity of phrases e.g. bigrams, trigrams. The TDNN exploits the property of parallelism allowing all feature maps to be computed in parallel yet obtaining competitive performance with an RNN encoder (Section 4.2). The phrase level representation obtained by the TDNN provides a trade-off between capturing semantics and computational time.

3.2 Decoder

Since the title of a manuscript is short but more concise, we require a finer grain representation than the phrase level of the TDNN. We adopt an RNN to represent the decoder with its large capacity to condition each word on all previous words in the sequence while considering its internal state and the encoder's representation. Let $\mathbf{x}_i^d$ be a e dimensional embedding corresponding to the i^{th} word of the cited document's title of length m. We utilize the Gated Recurrent Unit (GRU) [3] to help prevent the vanishing or exploding gradient problem, formally:

$$\begin{aligned}
\mathbf{z}_i &= \sigma(\mathbf{W}_z\mathbf{x}_i^d + \mathbf{V}_z\mathbf{c}_i + \mathbf{U}_z\mathbf{h}_{i-1}) \\
\mathbf{r}_i &= \sigma(\mathbf{W}_r\mathbf{x}_i^d + \mathbf{V}_r\mathbf{c}_i + \mathbf{U}_r\mathbf{h}_{i-1}) \\
\tilde{\mathbf{h}}_i &= \tanh(\mathbf{W}_o\mathbf{x}_i^d + \mathbf{V}_o\mathbf{c}_i + \mathbf{r}_i \circ \mathbf{U}_o\mathbf{h}_{i-1}) \\
\mathbf{h}_i &= (1 - \mathbf{z}_i) \circ \tilde{\mathbf{h}}_i + \mathbf{z}_i \circ \mathbf{h}_{i-1}
\end{aligned}$$

where $\mathbf{W}_{[z,r,o]}, \mathbf{V}_{[z,r,o]}, \mathbf{U}_{[z,r,o]}$ are weight matrices to be learned, $\tilde{\mathbf{h}}_i$ is the new updated hidden state, $\mathbf{z}_i$ is the update gate, $\mathbf{r}_i$ is the reset gate, $\sigma(\cdot)$ is the sigmoid function and $\circ$ is the element wise product.

Although the max pooling layer obtains the most relevant features present for a given filter, it treats each feature map with uniform importance and words on the margins of the sequence are neglected. The attention mechanism learns a weighted interpolation $\mathbf{c}_i$ dependent on all of the encoder's representation conditioned on previous decoder states obtaining a richer representation with:

$$\mathbf{c}_i = \sum_j \alpha_{ij}\mathbf{s}_j \qquad \text{where } \alpha_{ij} = \text{softmax}(\mathbf{v}^\top \tanh(\mathbf{W}_a\mathbf{h}_{i-1} + \mathbf{U}_a\mathbf{s}_j))$$

where α_{ij} is the alignment between the i^{th} word and the j^{th} output from the encoder parametrized as a feedforward neural network followed by a softmax function [5]. Figure 1 illustrates these recurrent dependencies with dashed arrows.

3.3 Author Networks

The author(s) of a manuscript may have a large impact on the audience, popularity, and citations. Frequently, one may follow specific researchers or groups with similar interests. The lead author of a paper may hold the most authority. On the other hand, the most influential author may not necessarily be the first author. To capture the most prominent author, we consider both the citing (context) and cited (title) manuscript authors with a shared embedding space, but learn two separate TDNNs. Intuitively, the author's characteristics may remain static hence the shared embedding space but the author has no direct control over if she will be cited or not (with the exception of self-citation). For example, a popular author may be frequently cited yet citations may not be reciprocated leading to distinct roles. We treat each author as a token by denoting $\mathbf{A}^q$ and $\mathbf{A}^d$ as the embeddings of the citation context (query) and cited paper's (document) author(s), respectively. Similar to the encoder representation presented in Section 3.1, we exploit the TDNN to learn higher level joint author interactions with:

$$\mathbf{s}_j = [f(\mathbf{X}^q) \oplus f(\mathbf{A}^q) \oplus f(\mathbf{A}^d)]_j \tag{2}$$

By concatenating the citation context summary with the author's representation, the attention mechanism conditions on the author networks in addition to the encoder's output. Hence an interaction between the composition of the context and author takes place over the course of the decoding process. The final output from the RNN decoder is projected into a softmax layer producing a probability over the vocabulary:

$$P(y_i|y_{\leq i}, \mathbf{s}) = \text{softmax}(\mathbf{V}\mathbf{h}_i)$$

where $P(y_i|y_{\leq i}, \mathbf{s})$ denotes the conditional probability of all previous words in the cited papers title prior to i. Since the entire architecture is differentiable, we jointly training the encoder-decoder via stochastic gradient descent (SGD) [5] maximizing the following:

$$\log P(\mathbf{y}|\mathbf{X^q}, \mathbf{X^d}, \mathbf{A^q}, \mathbf{A^d}) = \sum_i^m \log P(y_i|y_{\leq i}, \mathbf{s}) \tag{3}$$

Once the network is fully trained we can score a cited document $\mathbf{y}$ given a citation context $\mathbf{X}^q$ and author information $\mathbf{A}^q, \mathbf{A}^d$ with Equation 3.

	Recall	**MAP**	**MRR**	**NDCG**
BM-25	0.1007	0.0556	0.0606	0.0676
CTM	0.1288	0.0726	0.0777	0.0875
RNN-to-RNN	0.1590	0.0958	0.1054	0.1134
TDNN-to-RNN	0.1579	0.0935	0.1032	0.1114
Neural Citation Network	**0.2910**	**0.2418**	**0.2667**	**0.2592**

Table 1: Performance comparison of the top 10 recommendations on Recall, MAP, MRR, and NDCG. (NCN is statistically significant from all baselines on a paired t-test $p < 0.001$)

4 EXPERIMENTS

4.1 Setup

We evaluate NCN on the RefSeer dataset [2] [8]. After preprocessing invalid entries, we obtain 4,549,267 context pairs with 855,735 papers in a citation-cited relation. Similar to [8], we divide the data by year, where papers before, after, and equal to 2013 yield 4,258,383 training; 148,927 testing; and 141,957 validation citation contexts respectively. For text preprocessing, we perform tokenization, lemmatization and take the top 20K most frequent terms on the encoder and decoder sides, where words not on this list are replaced with a special <UNK> token. We also take top 20K most frequently cited authors by name and consider the first 5 authors per paper for simplicity. Authors not on the short list are replaced with a with a special <UNK>$_{\text{Author}}$.

All hyperparameters are determined according to the validation set. For clarity, we set all embedding sizes, batch sizes, RNN memory cell sizes and feature maps to 64. We apply gradient clipping at 5, dropout probability to 0.2 and the number of recurrent layers to two for both the encoder (when applicable) and decoder. For the NCN

[2] http://refseer.ist.psu.edu/data/

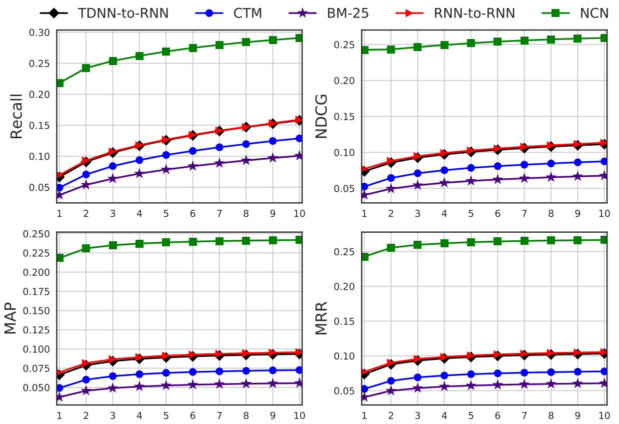

Figure 2: Recall, NDCG, MAP, and MRR as the number of recommendations vary from 1 to 10.

encoder, convolutional filters use region sizes: 4, 4, 5 and author networks use region sizes: 1, 2. We use the Adam optimizer [5] for a total of 5 training iterations, taking approximately 10 hours to train NCN on a NVIDIA Titan X.

We report the following metrics: Recall, Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and Normalized discounted cumulative gain (NDCG) on the test set. For NCN, we rerank the top 2048 documents retrieved by BM-25 with Equation 3 and include the ground truth if it is not present.

4.2 Baselines

We validate the effectiveness of NCN against four baselines: *BM-25*; *Citation Translation Model (CTM)* [7], we learn a translation model using the GIZA++ toolkit; *TDNN-to-RNN*, follows the NCN formulation excluding author networks; *RNN-to-RNN*, identical to TDNN-to-RNN but utilizing a RNN as the encoder.

Table 1 demonstrates NCN outperforms all baselines on every metric by 13-16%. BM-25 displays the poorest performance verifying the existence of the vocabulary gap while CTM[3] improves upon standard IR methods but the bag-of-words assumption lacks sufficient capacity to capture complex relations. Since NCN without author content degenerates to the TDNN-to-RNN model, we clearly see the advantages of incorporating author information. RNN-to-RNN marginally outperforms the TDNN-to-RNN model, however, the additional computational overhead may not justify the 0.3% increase in performance taking 11 hours to train yet NCN produces superior performance in less time. We observe smaller performance gains on position aware metrics in NCN when varying the number of recommendations. An improvement of 1.6% on NDCG, 2.4% on MAP and MRR when cutting off the number of recommendations at 10 versus 1 as illustrated in Figure 2.

4.3 Qualitative Study

The top three recommendations by NCN, CTM and RNN-to-RNN for the context (query) are listed in Table 2. Both baselines correctly recommend one item and NCN provides two correct recommendations; however, the incorrect recommendation (2) appears to be a plausible citation. We noticed the recommendations produced by NCN all have common authors[4]. Recommendations 1 and 3 contain M. Jordan as an author and recommendations 2 and 3 shares the author Z. Ghahramani further portraying NCNs successful integration of author information to produce relevant recommendations.

Context: "find a distribution over the latent variables that is close to the posterior of interest. Variational methods provide effective approximations in topic models and nonparametric Bayesian models"
Neural Citation Network
1. **Graphical models, exponential families, and variational inference** 2. Graphical models and variational methods 3. **An introduction to variational methods for graphical models**
CTM
1. Indexing by latent semantic analysis 2. **An introduction to variational methods for graphical models** 3. Bayesian data analysis
RNN-to-RNN
1. **An introduction to variational methods for graphical models** 2. The variational formulation of the Fokker–Planck equation 3. A Bayesian analysis of the multinomial probit model with fully identified parameters

Table 2: Top 3 recommendations for NCN, CTM and RNN-to-RNN for the citation context (query), correct recommendations are in bold.

5 CONCLUSIONS AND FUTURE WORK

We have introduced NCN, a flexible architecture capable of incorporating author metadata and highlight a promising new direction for context-aware citation recommendation. In future work, we plan to explore temporal aspects, and the large hyperparameters space such as filter strides, wide convolutions, dynamic k-max pooling and multi-channel convolutions.

REFERENCES

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[2] J. Beel, B. Gipp, S. Langer, and C. Breitinger. Research-paper recommender systems: a literature survey. *IJDL*, 17(4):305–338, 2016.

[3] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, 2014.

[4] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*, 2008.

[5] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[6] J. He, J.-Y. Nie, Y. Lu, and W. X. Zhao. Position-aligned translation model for citation recommendation. In *SPIRE*, 2012.

[7] W. Huang, S. Kataria, C. Caragea, P. Mitra, C. L. Giles, and L. Rokach. Recommending citations: Translating papers into references. In *CIKM*, 2012.

[8] W. Huang, Z. Wu, C. Liang, P. Mitra, and C. L. Giles. A neural probabilistic model for context based citation recommendation. In *AAAI*, 2015.

[9] Y. Lu, J. He, D. Shan, and H. Yan. Recommending citations with translation model. In *CIKM*, 2011.

[10] K. Xu, J. Ba, J. R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[3] Performance is less than reported in [8] due to significantly reduced vocabulary size.

[4] Authors omitted due to space constraints.