To Cite, or Not to Cite? Detecting Citation Contexts in Text

Michael Färber[1], Alexander Thiemann[1], and Adam Jatowt[2]

[1] University of Freiburg, Germany
michael.faerber@cs.uni-freiburg.de mail@athiemann.net
[2] Kyoto University, Japan
adam@dl.kuis.kyoto-u.ac.jp

Abstract. Recommending citations for scientific texts and other texts such as news articles has recently attracted considerable amount of attention. However, typically, the existing approaches for citation recommendation do not explicitly incorporate the question of whether a given context (e.g., a sentence), for which citations are to be recommended, actually "deserves" citations. Determining the "cite-worthiness" for each potential citation context as a step before the actual citation recommendation is beneficial, as (1) it can reduce the number of costly recommendation computations to a minimum, and (2) it can more closely approximate human-citing behavior, since neither too many nor too few recommendations are provided to the user. In this paper, we present a method based on a convolutional recurrent neural network for classifying potential citation contexts. Our experiments show that we can significantly outperform the baseline solution [1] and reduce the number of citation recommendations to about 1/10.

Keywords: Citation Context, Citation Recommendation, Recommender Systems, Deep Learning

1 Motivation

Due to a variety of reasons, such as supporting claims and arguments or giving attribution to authors, scientific works must contain appropriate citations to other works [2]. Citing properly is a challenging task: Not only all works leading to new results should be cited, but also adding citations to further explain concepts and ideas often helps the reader to correctly understand the goals and ideas of a paper. Finding a good balance between not too many and not too few citations is rather time consuming and requires years of practice in scientific writing. This issue also appears in the context of recommending citations automatically. Recommending citations for scientific texts and other texts such as news articles has recently attracted a considerable amount of attention, due to the dramatic increase in the number of published papers. Existing citation recommendation approaches, however, do not explicitly incorporate the question

of whether a given context (e.g., a sentence), for which citations should be recommended, actually "deserves" citations. If this is the case, we call it an actual *citation context*. In this paper, we approach this as a classification task and call it *citation context detection*. It can be regarded as a step before the actual recommendation of relevant citations. Note that the task of citation recommendation and, hence, citation context detection, is not limited to scientific texts such as publications, but can be applied to any text for which citations are needed, such as news texts or encyclopedic articles like those in Wikipedia.

Sugiyama et al. [1] provide the first approach to the presented research problem. In this paper, we show that the problem can be solved more effectively by a convolutional recurrent neural network. Our experiments reveal the superiority of our approach and offer insights into human-citing behavior.

2 Related Work

Citation context characterization and classification. Explicitly classifying potential citation contexts with respect to cite-worthiness has been carried out by Sugiyama et al. [1] by means of an SVM approach. However, they only report the accuracy of results and do not address the high imbalance of negative to positive instances. In [3], the authors focus on the distributions of citation locations in publications, although they only provide visual analyses and no prediction model. Angrosh et al. [4] only consider sentences in related work sections and classify them into 13 classes. We cannot apply their classification scheme or approach, as we do not only consider related work sections, so our sentences are of a different nature. Citation contexts have also been studied in further respects. Most prominently, the citation function has been analyzed and predicted [5], and the citation importance [6] and further linguistic characteristics such as the discourse structure of citation contexts [7] were also analyzed.

Citation recommendation. For citation recommendation, a variety of approaches have been proposed (see [8], but we note that most listed methods approach the paper recommendation task, which differs from the citation recommendation by using information from the entire paper instead of short citation contexts only). Most recent approaches typically utilize neural networks and learning-to-rank-frameworks [9,10,11]. Typically, the citation contexts are already predetermined. Using our approach, such prerequisites are no longer necessary. We believe that having a flexible approach that determines the placement of citations themselves is more user-friendly and can be applied in a variety of scenarios. Furthermore, in this way we can reduce the number of citation recommendations, as we only focus on the cite-worthy contexts for the recommendation.

3 Citation Context Classification

We now describe our approach to identify citation contexts. As potential citation contexts, we use single sentences since sentences are a natural unit for expressing

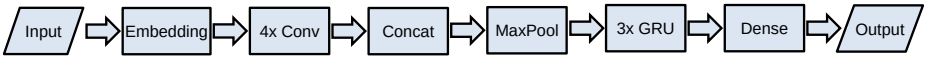

Fig. 1: Our network used for citation context detection.

statements and since prior studies have revealed that there is no single optimal choice for a citation context unit (such as sentences or a fixed window) [12]. Our method attempts to classify sentences into "needing citations" and "not needing citations." As an underlying method, we use a convolutional recurrent neural network (CRNN) [13], as it was shown to be a good fit for text classification. A pure recurrent neural network (RNN) classifier is biased: Later words dominate compared to earlier words. This is unnatural for documents, as important information could be spread in the text. This problem can be addressed with convoluational neural networks (CNNs), but picking a good window size is challenging, as it could lead to the loss of important information. Thus, we combine both methods to obtain state-of-the-art performance.

The full architecture of the CRNN[3] – visible in Fig. 1 – consists of four convolutional layers with 128 hidden states with a filter size of 1, 2, 3, and 5. Next is a concatenation step followed by max pooling. After the convolutional part, the recurrent part consists of three gated recurrent unit (GRU) [14] layers with a (recurrent) dropout of 0.2. Finally, a densely connected layer with a softmax activation function and two outputs provides the final classification.

4 Evaluation

4.1 Evaluation Data Sets

For evaluating our approach, the contents of publications are needed. Note that many available scholarly data sets either only cover the citation contexts and not all sentences of the publications (e.g., see CiteSeerX) or only cover meta-information about the publications, such as the citation network. After reviewing scholarly data sets, we decided on using the following:[4]

arXiv CS [15] is a data set of over 9M sentences extracted from all computer science publications hosted at arXiv.org. This data set was constructed from the TEX files provided by the authors. Since each citation is explicitly given via a `cite` command in TEX, for this data set we can ensure that we do not miss any citations and that we always link to the correct reference. This makes this data set to be of relatively high quality.

Scholarly Dataset 2[5] contains about 100k publications in PDF format from the ACM Digital Library. By using this data set, we can evaluate the impact of having PDFs as input, which, in reality, is often the case.

[3] The source code is available online at `https://github.com/agrafix/grabcite-net`.
[4] All data sets are available online at `http://citation-recommendation.org/publications`.
[5] `http://www.comp.nus.edu.sg/~sugiyama/SchPaperRecData.html`.

ACL-ARC[6] is a widely used corpus of scholarly publications about computational linguistics. In order to compare against Sugiyama et al.'s approach [1], we use the 2008 version, which contains 10,921 papers.

For transforming the PDF files of Scholarly and ACL-ARC into plaintext files, we use IceCite [16], which is a state-of-the-art information extraction tool for scientific publications.

4.2 Evaluation of Citation Context Detection

Training data. We build training data by iterating over all sentences in our input data (i.e., plaintext of publications), detecting if a citation marker is present, and labeling the sentences accordingly before removing the citation markers from the sentences. This gives us heavily imbalanced training data: 10% of all sentences contain at least one reference and 90% do not. We solve this imbalance by oversampling for all NNs and by undersampling for all SVM approaches. We use the pretrained GloVe word embeddings ("GloVe 6B") for our NNs. For arXiv CS (since it is the cleanest data set), we also try our own word embeddings trained via auto encoding.

Methods. We use the following approaches for a comparison:

- **SVM**. Following Sugiyama et al. [1], we use a Support Vector Machine (SVM) and try out different feature settings listed in Table 3.
- **CNN**. Secondly, we try a convolutional neural network (CNN).
- **RNN**. Then, a recurrent neural network (RNN) is also used.
- **CRNN**. This is our approach proposed in Section 3 using five epochs with a batch size of 64.

Evaluation results. The results are given in Table 1. When considering accuracy as metric, already simple approaches like $\text{SVM}_{\text{doc2vec}}$ perform very well and outperform NNs. However, accuracy is not a very suitable metric in the case of unbalanced data (as given here); for instance, using a SVM with doc2vec as the only feature achieved very good accuracy scores, too, but low F1 scores.

Sugiyama's approach [1], which is based on a SVM with noun phrases or the indication about citations in the neighboring sentences as a feature, can be outperformed by many of our approaches. The better accuracy value of our SVM_{NPs} for ACL-ARC compared to $\text{SVM}_{[1]}$ may be a result from improved PDF-to-text conversion.

Considering the F1 scores, our NNs outperform the SVM approaches. Interestingly, both the CNN and the RNN perform roughly the same as the CRNN. This indicates that for citation context detection, the ordering of the words is not that important and that the word embeddings themselves already have a strong signal for the classification. This is plausible if we consider that named entities and abstract concepts rather than complete statements are cited.

[6] `http://acl-arc.comp.nus.edu.sg/`.

Table 1: Results of classifying sentences regarding their cite-worthiness (P: Precision, R: Recall, F1: F1 score, A: Accuracy).

	ACL-ARC				Scholarly				arXiv CS			
	P	R	F1	A	P	R	F1	A	P	R	F1	A
SVM of [1]	-	-	-	0.882	-	-	-	-	-	-	-	-
$SVM_{TF\text{-}IDF}$	0.049	0.052	0.051	0.938	0.043	0.009	0.015	0.969	0.100	**0.955**	0.180	0.131
SVM_{POS}	0.061	**0.658**	0.112	0.670	0.050	0.680	0.093	0.662	0.191	0.631	0.293	0.695
SVM_{NPs}	0.034	0.050	0.041	0.926	0.019	0.001	0.002	**0.973**	0.094	0.048	0.063	0.858
$SVM_{doc2vec}$	0.086	0.004	0.008	**0.967**	0.026	**1.000**	0.050	0.028	0.140	0.016	0.029	**0.892**
SVM_{PER}	0.049	0.099	0.066	0.912	0.116	0.137	0.126	0.951	0.338	0.186	0.240	0.882
SVM_{Cits}	0.083	0.578	0.145	0.786	0.068	0.509	0.120	0.809	0.199	0.724	0.313	0.681
CNN_{GloVe}	**0.196**	0.269	**0.227**	0.941	**0.227**	0.792	**0.329**	0.812	**0.433**	0.709	**0.538**	0.870
RNN_{GloVe}	0.171	0.317	0.222	0.928	0.181	0.823	0.322	0.811	0.400	0.785	0.530	0.851
$CRNN_{GloVe}$	0.182	0.260	0.214	0.930	0.207	0.763	0.326	0.807	0.376	0.750	0.501	0.841

Table 2: Results for arXiv data set using custom word embeddings instead of pre-trained GloVe word embeddings.

	P	R	F1	A
CNN_{custom}	0.418	0.724	0.530	0.863
RNN_{custom}	0.393	**0.790**	0.525	0.849
$CRNN_{custom}$	**0.430**	0.715	**0.537**	**0.869**

Table 3: Features used for our SVM approach.

Name	Description
$SVM_{TF\text{-}IDF}$	TF-IDF
SVM_{POS}	# POS tags
SVM_{NPs}	BOW of extracted noun phrases
$SVM_{doc2vec}$	doc2vec
SVM_{PER}	contains person accord. to Stanford NER
SVM_{Cits}	# citations in neighb. sentences (up to 5)

We can observe significant differences in the evaluation scores for all approaches between the different data sets. The reason is likely to be the varying quality of the data sets. The contents of ACL-ARC and Scholarly are extracted from PDFs and are thus quite noisy (especially ACL-ARC), while arXiv CS remains comparatively clean. Thus, we can assume that arXiv CS best reflects the actual citing behavior. For this data set (and for other data sets), the results of the NN approaches only vary very little. If instead of pretrained GloVe word embeddings, our own embeddings are trained – as done exemplarily for the arXiv data set – we achieve considerably better precision and F1 scores for the CRNN (see Table 2), while the results for the CNN slightly decrease and remain stable for the RNN. Hence, under the assumption of having a larger training corpus in the future, the $CRNN_{custom}$ seems to be one of the most promising approaches. In total, we achieve F1 scores of over 0.5 for all NNs on the arXiv CS data set, making citation context detection attractive to be applied in actual systems.

5 Conclusion and Outlook

Existing citation recommendation approaches do not incorporate the question of whether a given citation context actually deserves citations. In this paper, we address this question and build a classifier that can determine the

"cite-worthiness" to a considerable degree. As a result, we can reduce the number of costly citation recommendation computations to a minimum (about 1/10), since recommendations need to be computed only for cite-worthy contexts and since only about 10% of the sentences in our data sets contain citations. Our experiments on three data sets show that we can significantly outperform the existing solution of Sugiyama et al. [1]. For future work, we plan to consider additional features using the papers' meta-data.

Acknowledgements. Michael Färber is an International Research Fellow of the Japan Society for the Promotion of Science (JSPS). The work was partially supported by MIC SCOPE (171507010). The Titan Xp used for this research was donated by the NVIDIA Corporation.

References

1. Sugiyama, K., Kumar, T., Kan, M.Y., Tripathi, R.C.: Identifying Citing Sentences in Research Papers Using Supervised Learning. CAMP 2010, IEEE (2010) 67–72
2. Teufel, S., Siddharthan, A., Tidhar, D.: An annotation scheme for citation function. SIGdial '09 (2009) 80–87
3. Hu, Z., Chen, C., Liu, Z.: Where are citations located in the body of scientific articles? A study of the distributions of citation locations. J. Informetrics **7**(4) (2013) 887–896
4. Angrosh, M.A., Cranefield, S., Stanger, N.: Context Identification of Sentences in Related Work Sections using a Conditional Random Field: Towards Intelligent Digital Libraries. JCDL 2010 (2010) 293–302
5. Teufel, S., Siddharthan, A., Tidhar, D.: Automatic classification of citation function. EMNLP 2007 (2006) 103–110
6. Valenzuela, M., Ha, V., Etzioni, O.: Identifying Meaningful Citations. SBD'15 (2015)
7. Fisas, B., Saggion, H., Ronzano, F.: On the Discoursive Structure of Computer Graphics Research Papers. In: LAW@NAACL-HLT 2015. (2015) 42–51
8. Beel, J., Gipp, B., Langer, S., Breitinger, C.: Research-paper recommender systems: a literature survey. Int. J. on Digital Libraries **17**(4) (2016) 305–338
9. Ebesu, T., Fang, Y.: Neural Citation Network for Context-Aware Citation Recommendation. SIGIR'17 (2017) 1093–1096
10. Jiang, Z., Liu, X., Gao, L.: Chronological Citation Recommendation with Information-Need Shifting. CIKM'15 (2015) 1291–1300
11. Huang, W., Wu, Z., Chen, L., Mitra, P., Giles, C.L.: A Neural Probabilistic Model for Context Based Citation Recommendation. AAAI'15 (2015) 2404–2410
12. Alvarez, M.H., Gómez, J.M.: Survey about citation context analysis: Tasks, techniques, and resources. Natural Lang. Eng. **22**(3) (2016) 327–349
13. Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent Convolutional Neural Networks for Text Classification. AAAI'15 (2015) 2267–2273
14. Zhou, G., Wu, J., Zhang, C., Zhou, Z.: Minimal Gated Unit for Recurrent Neural Networks. CoRR **abs/1603.09420** (2016)
15. Färber, M., Thiemann, A., Jatowt, A.: A High-Quality Gold Standard for Citation-based Tasks. LREC 2018 (2018)
16. Bast, H., Korzen, C.: A benchmark and evaluation for text extraction from PDF. JCDL 2017 (2017) 99–108