CITEWERTs: A System Combining Cite-Worthiness with Citation Recommendation

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. Due to the vast amount of publications appearing in the various scientific disciplines, there is a need for automatically recommending citations for text segments of scientific documents. Surprisingly, only few demonstrations of citation-based recommender systems have been proposed so far. Moreover, existing solutions either do not consider the raw textual context or they recommend citations for predefined citation contexts or just for whole documents. In contrast to them, we propose a novel two-step architecture: First, given some input text, our system determines for each potential citation context, which is typically a sentence long, if it is actually "cite-worthy." When this is the case, secondly, our system recommends citations for that context. Given this architecture, in our demonstration we show how we can guide the user to only those sentences that deserve citations and how to present recommended citations for single sentences. In this way, we reduce the user's need to review too many sentences and recommendations.

Keywords: Citation Recommendation, Citation Context, Digital Libraries, Recommender Systems

1 Motivation

The number of published papers within the different scientific disciplines has increased dramatically in the last years: More than 100,000 new computer science papers are published every year [1]. A similar trend can be observed in other disciplines. This overload of information leads to the fact that scientists (and other people who need to cite facts, such as authors of news articles and editors of encyclopedias) cannot be aware of all publications at any given time and hence have difficulties during citing. To assist users in the process of citing, citation recommendation approaches have been proposed that recommend citations based on a given text fragment, which is called the *citation context* and which can be a sentence in a paper. However, most approaches (such as [2,3,4], to name only the most recent) do not explicitly incorporate the question of whether a given citation context for which citations are to be recommended,

actually requires citations. As a consequence, citations are recommended for each potential citation context (e.g., sentence), even though the context may not "need" any citation in the first place.

This also holds in the case of presented demonstrations of citation recommendation systems. The RefSeer system [5], which is the most related demonstration to ours (though apparently no longer available online), recommends one citation for each sentence in the input text. Besides [5], to the best of our knowledge, only paper recommendation systems exist, i.e., systems that do not use any citation context, but, for instance, just use a citation graph [6]. *TheAdvisor* [7] and *FairScholar* [8] are further examples of paper recommender system demonstrations.

In this paper, we demonstrate an end-to-end system that is, to the best of our knowledge, the first one to combine two steps that have been considered separately so far, namely (1) determining the cite-worthiness of every potential citation context (e.g., sentence) and (2) recommending citations for "approved" citation contexts. As a consequence, our system does not require knowing a priori the locations in which a citation should be inserted; instead, they will be determined automatically (up to the sentence level). This two-step approach is not only more user-friendly, as it hides unnecessary recommendations, but it also reduces the number of costly recommendation computations.

The user of our system, which can be any scientist who is writing or reading text, can use the front-end to be guided to sentences that apparently need citations; he can then review the recommended citations for the single citation contexts. By training our approach on the publications hosted at arXiv.org (until August 2017), we are able to recommend papers published until recently. The source code of our system is available online.[3]

2 System Overview

Before describing the user interface and a typical workflow of the system, we outline the back-end. This consists of two main components:

Sentence Classification. This component makes a binary classification whether the given potential citation context needs a citation or not. We implement this classification step by means of a convolutional recurrent neural network (CRNN). The full architecture of the CRNN consists of four convolutional layers with 128 hidden states and with filter sizes 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) 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. An evaluation of this classification approach is provided in [9].

Citation Recommendation. By manually evaluating 1,500 randomly chosen sentences w.r.t. the characteristics of the citation contexts and w.r.t.

[3] See `https://github.com/agrafix/grabcite` and `https://github.com/agrafix/grabcite-net`.

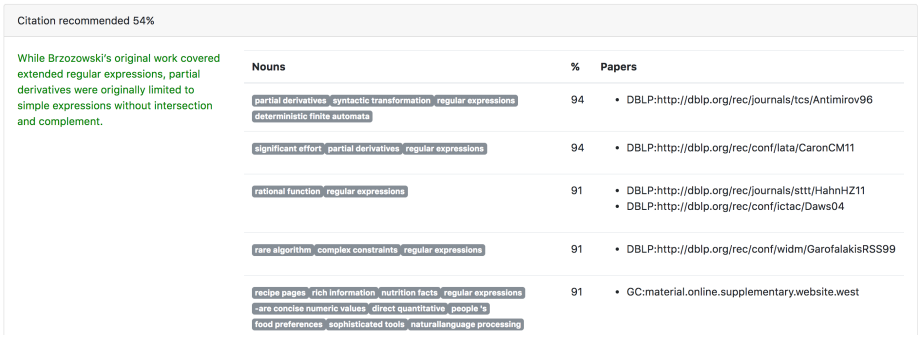

Fig. 1: Our system's user interface after selecting a cite-worthy sentence.

their cite-worthiness,[4] we discover that citations mainly relate to a specific noun phrase mentioned in the citation context. This noun phrase is often either the name(s) of the cited paper's author(s) (e.g., "Carignan et al. [X] have"), or it is a specifically named technique or method (e.g., "discrete exterior calculus (DEC) [X]"), or an abstract concept (e.g., "recurrent neural networks"). Based on these insights, our idea is to extract all nouns from a sentence and to train a latent semantic index (LSI) [11] via TF-IDF on our data set.[5] Since LSI is only based on the citation contexts of the citing papers, no content of the *cited* papers is needed, making our approach applicable in many scenarios. To build the LSI, we only consider sentences that contain at least one citation.

We use the Python library TextBlob to extract the noun phrases. After splitting the data set into a training set (90%) and a test set (10%), the LSI model is computed with n=200 factors via a fast truncated SVD [12]. In the testing phase, we convert the sequence of extracted nouns into LSI vector space and query the LSI index by computing similarities. Finally, we use a metadata mapping to recover contained references to the most similar entries of the LSI index, which are then returned as recommendations to the user.

3 Interface Usage

The interface of our system is available online at `http://citewerts.citation-recommendation.org`. A prototypical user inputs some text (e.g., a text paragraph without citations) into the input text box and presses "Analyze." Instantly, this input is processed in the background: Internally, the text is split into potential citation contexts (sentences). Each context is then classified as cite-worthy or not. In our interface, we then show all sentences and highlight the

[4] We use the documents of the DRI corpus [10], as they have already been used for other citation analysis studies. The assessments, conducted by the authors, are available online at `http://citation-recommendation.org/publications`.

[5] Note that our intention was not to build a novel approach for citation recommendation that outperforms existing methods but to show the usefulness of combining citation recommendation with citation context classification.

cite-worthy sentences (the darker they are, the more confident the network was). When a user clicks on a sentence, the citation recommendations obtained via the trained LSI model are shown. For each recommended citation, the terms shown during training for this citation and the matching score (in percent) are also displayed. Although the visualization is currently restricted to the publication identifiers in our index (using DBLP URIs, if the publication is in DBLP), it can easily be extended to show the meta-information directly.

4 Conclusions

In this paper, we demonstrated a system that not only recommends citations but firstly identifies cite-worthy contexts in the input text. Our system is user-friendly, since it hides unnecessary recommendations, and it reduces the number of costly recommendation computations. It can be used not only for scientific texts in different disciplines but also, for instance, for texts from encyclopedias or news articles.

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).

References

1. Küçüktunç, O., Saule, E., Kaya, K., Çatalyürek, Ü.V.: Diversifying Citation Recommendations. ACM Trans. Intell. Syst. Technol. **5**(4) (2014) 55:1–55:21
2. Ebesu, T., Fang, Y.: Neural Citation Network for Context-Aware Citation Recommendation. SIGIR'17 (2017) 1093–1096
3. Jiang, Z., Liu, X., Gao, L.: Chronological Citation Recommendation with Information-Need Shifting. CIKM'15 (2015) 1291–1300
4. 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
5. Huang, W., Wu, Z., Mitra, P., Giles, C.L.: RefSeer: A citation recommendation system. In: JCDL 2014. (2014) 371–374
6. Huynh, T., Hoang, K., Do, L., Tran, H., Luong, H.P., Gauch, S.: Scientific Publication Recommendations Based on Collaborative Citation Networks. CTS'12 (2012) 316–321
7. Küçüktunç, O., Saule, E., Kaya, K., Çatalyürek, Ü.V.: TheAdvisor: A Webservice for Academic Recommendation. JCDL '13 (2013) 433–434
8. Anand, A., Chakraborty, T., Das, A.: FairScholar: Balancing Relevance and Diversity for Scientific Paper Recommendation. ECIR 2017 (2017) 753–757
9. Färber, M., Thiemann, A., Jatowt, A.: To Cite, or Not to Cite? Detecting Citation Contexts in Text. ECIR 2018 (2018)
10. Fisas, B., Ronzano, F., Saggion, H.: A Multi-Layered Annotated Corpus of Scientific Papers. LREC'16 (2016)
11. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by Latent Semantic Analysis. JASIS **41**(6) (1990) 391–407
12. Boutsidis, C., Magdon-Ismail, M.: Faster SVD-Truncated Least-Squares Regression. CoRR **abs/1401.0417** (2014)