BibTeX-based dataset generation for training citation parsers

Sree Harsha Ramesh, Dung Thai,
Boris Veytsman, Andrew McCallum

A citation graph is an important part of modern scientometrics (the field of analyzing and measuring of scientific literature) [2–19,21,23–31]. To construct it, we need to disambiguate citations: determine which paper cites which paper. While many publishers now deposit citation data in a machine readable format, some do not — and there are millions of older papers where only textual citation strings are available. Since manual conversion of these strings to parsed entries is not possible, we need to teach machines how to do this.

An important part of supervised learning is a good dataset of *ground truth* — in our case, a large amount of already parsed citations both as text strings and key-value pairs. The traditional way to generate these datasets is to take a large number of citations and manually parse each of them. This process is tedious and expensive, since in many cases it requires trained annotators. Therefore the existing datasets are relatively small: the CORA Field Extraction dataset [22] has 500 citations, and the UMass Citation Field Extraction dataset [1] has 1829 citations.

Our new approach to creating the dataset overcomes this difficulty. We start with already parsed data: BibTeX files of papers. Using different bibliography styles (`bst` files), we generate formatted citations, for which we know the content in the key-value format as we used this content to create the formatted text.

Initially we intended to use Nelson Beebe's extensive BibTeX archives.[1] However, we discovered that the bibliographies there are not suitable for our task: they have a large, but still limited number of journals, they do not have "unusual" fields like `eprint`, and they do not have the errors and inconsistencies often encountered in the wild. Therefore software trained on Beebe's files were not very successful in parsing "wild" citations.

So, we used another approach. We scraped the Internet for `.bib` files, finding 9393 BibTeX files (mostly personal bibliographies) with 1 216 607 entries. We manually cleaned them, deleting duplicate fields, missing delimiters, unenclosed braces, etc. We used 297 `bst` files from TeX Live. The resulting dataset is described in Table 1. The size of this dataset is several orders of magnitude larger than the largest previously available [1].

We trained a number of modern algorithms for citation parsing based on our dataset. The results for the ELMO tagger [20] are shown in Tables 2 and 3 with the common accuracy measure $F1$ (the harmonic mean of recall and precision) shown.

It is interesting to see how use of the BibTeX dataset improves the performance of the tagger, as trained and tested on the UMass dataset [1]. The results are shown in Table 4. We see a significant

[1] `http://math.utah.edu/~beebe/bibliographies.html`

Table 1: Generated dataset

Parameter	Value
Total number of annotated citations	353 892 568
Vocabulary size	179 682
Total number of styles	237
Total number of field types	55
Total number of BibTeX source files	9393

Table 2: Field extraction performance on a subset of data (ELMO tagger)

Best fields Label	F1	*Worst fields* Label	F1
Ref-marker	99.99	Type	86.64
CODEN	99.74	E-Print	85.71
Year	99.73	Issue	80.00
ISSN	99.72	Price	80.00
Pages	99.63	How-Published	75.15
Volume	99.33	Organization	69.95
Number	99.32	Key	60.59
DOI	99.32	EID	54.84
Language	99.31	Comment	40.00
Month	99.25	Annote	30.77

Table 3: Performance for different BibTeX styles

Style	Recall	Precision	F1
The styles with the highest scores			
`swealpha`	98.21	99.00	98.60
`unsrtnat`	98.51	99.02	98.76
`ACM-Reference`	97.24	97.66	97.45
The styles with the lowest scores			
`ksfh_nat`	94.74	95.66	95.19
`rsc`	95.34	96.45	95.89
`gp`	95.60	96.37	95.98

Table 4: Improvement in UMass dataset parsing

Training	Recall	Precision	F1
UMass	93.58	94.02	93.80
BibTeX	94.25	93.18	93.78
UMass + BibTeX	97.59	97.23	**97.41**

improvement in the parsing of the existing dataset when additional data are added for training.

In conclusion, programmable typesetting and formatting systems like TeX and BibTeX can create "natural" text from structured data. This pseudo-natural text can be used to train machines.

◇ Sree Harsha Ramesh
College of Information and Computer Sciences, UMass Amherst
shramesh (at) cs dot umass dot edu

◇ Dung Thai
College of Information and Computer Sciences, UMass Amherst
dthai (at) cs dot umass dot edu

◇ Boris Veytsman
Meta, Chan Zuckerberg Initiative
bveytsman (at) chanzuckerberg dot com

◇ Andrew McCallum
College of Information and Computer Sciences, UMass Amherst
mccallum (at) cs dot umass dot edu

References

[1] S. Anzaroot and A. McCallum. A new dataset for fine-grained citation field extraction. In *Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA*, 2013.

[2] L. Bornmann, K. B. Wray, and R. Haunschild. Citation concept analysis (CCA)—a new form of citation analysis revealing the usefulness of concepts for other researchers illustrated by two exemplary case studies including classic books by Thomas S. Kuhn and Karl R. Popper. *arXiv e-prints* 1905.12410, May 2019.

[3] C. Castillo, D. Donato, and A. Gionis. Estimating number of citations using author reputation. In N. Ziviani and R. Baeza-Yates, eds., *String Processing and Information Retrieval: 14th International Symposium, SPIRE 2007 Santiago, Chile, October 29-31, 2007 Proceedings*, pp. 107–117. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. doi:10.1007/978-3-540-75530-2_10

[4] T. Chakraborty, S. Kumar, et al. Towards a stratified learning approach to predict future citation counts. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '14, pp. 351–360, Piscataway, NJ, USA, 2014. IEEE Press. http://dl.acm.org/citation.cfm?id=2740769.2740830

[5] C. Chen. Predictive effects of structural variation on citation counts. *Journal of the American Society for Information Science and Technology* 63(3):431–449, 2012. doi:10.1002/asi.21694

[6] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In *Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR*, 2007. https://icml.cc/imls/conferences/2007/proceedings/papers/257.pdf

[7] S. Feldman, K. Lo, and W. Ammar. Citation count analysis for papers with preprints. *ArXiv e-prints* 1805.05238, May 2018.

[8] L. D. Fu and C. F. Aliferis. Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics* 85(1):257–270, 2010. doi:10.1007/s11192-010-0160-5

[9] D. Herrmannova, P. Knoth, and R. Patton. Analyzing citation-distance networks for evaluating publication impact. In *11th edition of the Language Resources and Evaluation Conference*, May 2018. http://oro.open.ac.uk/53638/

[10] D. Herrmannova, R. M. Patton, et al. Do citations and readership identify seminal publications? *CoRR* abs/1802.04853, 2018. http://arxiv.org/abs/1802.04853

[11] B. I. Hutchins, X. Yuan, et al. Relative citation ratio (RCR): A new metric that uses citation rates to measure influence at the article level. *PLOS Biology* 14(9):1–25, 09 2016. doi:10.1371/journal.pbio.1002541

[12] I. Iacopini, S. Milojević, and V. Latora. Network dynamics of innovation processes. *Phys. Rev. Lett.* 120:048301, Jan 2018. doi:10.1103/PhysRevLett.120.048301

[13] Y. Jia and L. Qu. Improve the performance of link prediction methods in citation network by using h-index. In *2016 International Conference on Cyber-Enabled Distributed*

Computing and Knowledge Discovery (CyberC), pp. 220–223, Oct 2016. doi:10.1109/CyberC.2016.51

[14] M. Kaya, M. Jawed, et al. Unsupervised link prediction based on time frames in weighted–directed citation networks. In R. Missaoui, T. Abdessalem, and M. Latapy, eds., *Trends in Social Network Analysis: Information Propagation, User Behavior Modeling, Forecasting, and Vulnerability Assessment*, pp. 189–205. Springer International Publishing, Cham, 2017. doi:10.1007/978-3-319-53420-6_8

[15] P. Klimek, A. S. Jovanovic, et al. Successful fish go with the flow: citation impact prediction based on centrality measures for term–document networks. *Scientometrics* 107(3):1265–1282, Jun 2016. doi:10.1007/s11192-016-1926-1

[16] C. Lokker, K. A. McKibbon, et al. Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study. *BMJ* 336(7645):655–657, 2008. doi:10.1136/bmj.39482.526713.BE

[17] K. McKeown, I. Daume, Hal, et al. Predicting the impact of scientific concepts using full-text features. *Journal of the Association for Information Science and Technology* 67(11):2684–2696, 2016. doi:10.1002/asi.23612

[18] H.-M. Park, Y. B. Sinshaw, and K.-A. Sohn. Temporal citation network-based feature extraction for cited count prediction. In K. J. Kim and N. Joukov, eds., *Mobile and Wireless Technologies 2017: ICMWT 2017*, pp. 380–388. Springer Singapore, Singapore, 2018. doi:10.1007/978-981-10-5281-1_41

[19] B. K. Peoples, S. R. Midway, et al. Twitter predicts citation rates of ecological research. *PLoS ONE* 11:e0166570, 2017. doi:0.1371/journal.pone.0166570

[20] M. E. Peters, M. Neumann, et al. Deep contextualized word representations. In *NAACL*, 2018.

[21] N. Pobiedina and R. Ichise. Citation count prediction as a link prediction problem. *Applied Intelligence* 44(2):252–268, Mar 2016. doi:10.1007/s10489-015-0657-y

[22] K. Seymore, A. McCallum, and R. Rosenfeld. Learning hidden Markov model structure for information extraction. In *AAAI-99 workshop on machine learning for information extraction*, pp. 37–42, 1999.

[23] X. Shi, J. Leskovec, and D. A. McFarland. Citing for high impact. *CoRR* abs/1004.3351, 2010. http://arxiv.org/abs/1004.3351

[24] H. Small, K. W. Boyack, and R. Klavans. Citations and certainty: a new interpretation of citation counts. *Scientometrics* 118(3):1079–1092, Mar 2019. doi:10.1007/s11192-019-03016-z

[25] I. Tahamtan, A. Safipour Afshar, and K. Ahamdzadeh. Factors affecting number of citations: A comprehensive review of the literature. *Scientometrics* 107(3):1195–1225, June 2016. doi:10.1007/s11192-016-1889-2

[26] I. Tahamtan, A. Safipour Afshar, and K. Ahamdzadeh. Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics* 107(3):1195–1225, Jun 2016. doi:10.1007/s11192-016-1889-2

[27] B. Veytsman. How to measure the consistency of the tagging of scientific papers? In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 372–373, 2019. doi:10.1109/JCDL.2019.00076

[28] L. Weihs and O. Etzioni. Learning to predict citation-based impact measures. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 1–10, June 2017. doi:10.1109/JCDL.2017.7991559

[29] R. Yan, C. Huang, et al. To better stand on the shoulder of giants. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '12, pp. 51–60, New York, NY, USA, 2012. ACM. doi:10.1145/2232817.2232831

[30] R. Yan, J. Tang, et al. Citation count prediction: learning to estimate future citations for literature. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pp. 1247–1252, New York, NY, USA, 2011. ACM. doi:10.1145/2063576.2063757

[31] T. Yu, G. Yu, et al. Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics* 101(2):1233–1252, Nov. 2014. doi:10.1007/s11192-014-1279-6