Synthetic vs. Real Reference Strings for Citation Parsing, and the Importance of Re-training and Out-Of-Sample Data for Meaningful Evaluations: Experiments with GROBID, GIANT and CORA

Mark Grennan
Trinity College Dublin, School of Computer Science and Statistics, Ireland
grennama@tcd.ie

Joeran Beel
University of Siegen, Department of Electrical Engineering and Computer Science, Germany
joeran.beel@uni-siegen.de

Abstract

Citation parsing, particularly with deep neural networks, suffers from a lack of training data as available datasets typically contain only a few thousand training instances. Manually labelling citation strings is very time-consuming, hence, synthetically created training data could be a solution. However, as of now, it is unknown if synthetically created reference-strings are suitable to train machine learning algorithms for citation parsing. To find out, we train Grobid, which uses Conditional Random Fields, with a) human-labelled reference strings from 'real' bibliographies and b) synthetically created reference strings from the GIANT dataset. We find[1] that both synthetic and organic reference strings are equally suited for training Grobid (F1 = 0.74). We additionally find that retraining Grobid has a notable impact on its performance, for both synthetic and real data (+30% in F1). Having as many types of labelled fields as possible during training also improves effectiveness, even if these fields are not available in the evaluation data (+13.5% F1). We conclude that synthetic data is suitable for training (deep) citation parsing models. We further suggest that in future evaluations of reference parsing tools, both evaluation data being similar and data being dissimilar to the training data should be used to obtain more meaningful results.

1 Introduction

Accurate citation data is needed by publishers, academic search engines, citation & research-paper recommender systems and others to calculate impact metrics (Nisa Bakkalbasi et al., 2006; Jacso, 2008), rank search results (Beel and Gipp, 2009a,b), generate recommendations (Beel et al., 2016; Eto, 2019; Färber et al., 2018; Färber and Jatowt, 2020; Jia and Saule, 2018; Livne et al., 2014) and other applications e.g. in the field of bibliometric-enhanced information retrieval (Cabanac et al., 2020). Citation data is often parsed from unstructured bibliographies found in PDF files on the Web (Figure 1). To facilitate the parsing process, a dozen (Tkaczyk et al., 2018a) open source tools were developed including ParsCit (Councill et al., 2008), Grobid (Lopez, 2009, 2013), and Cermine (Tkaczyk et al., 2015). Grobid is typically considered the most effective one (Tkaczyk et al., 2018a). There is ongoing research that continuously leads to novel citation-parsing algorithms including deep learning algorithms (An et al., 2017; Bhardwaj et al., 2017; Nasar et al., 2018; Prasad et al., 2018; Rizvi et al., 2019; Rodrigues Alves et al., 2018; Zhang, 2018) and meta-learned ensembles (Tkaczyk et al., 2018c,b).

Most parsing tools apply supervised machine learning (Tkaczyk et al., 2018a) and require labelled training data. However, training data is rare compared to other disciplines where datasets may have millions of instances. To the best of our knowledge, existing citation-parsing datasets typically contain a few thousand instances and are domain specific (Figure 2). This may be sufficient for traditional machine learning algorithms but not for deep learning, which shows a lot of potential for citation parsing (An et al., 2017; Bhardwaj et al., 2017; Nasar et al., 2018; Prasad et al., 2018; Rizvi et al., 2019). Even for traditional machine learning, existing datasets may not be ideal as they often lack diversity in terms of citation styles.

Recently, we published GIANT, a synthetic dataset with nearly 1 billion annotated reference strings (Grennan et al., 2019). More precisely, the dataset contains 677,000 unique reference strings, each in around 1,500 citation styles (e.g. APA,

[1] The work presented in this manuscript is based on Mark Grennan's Master thesis "1 Billion Citation Dataset and Deep Learning Citation Extraction" at Trinity College Dublin, Ireland, 2018/2019

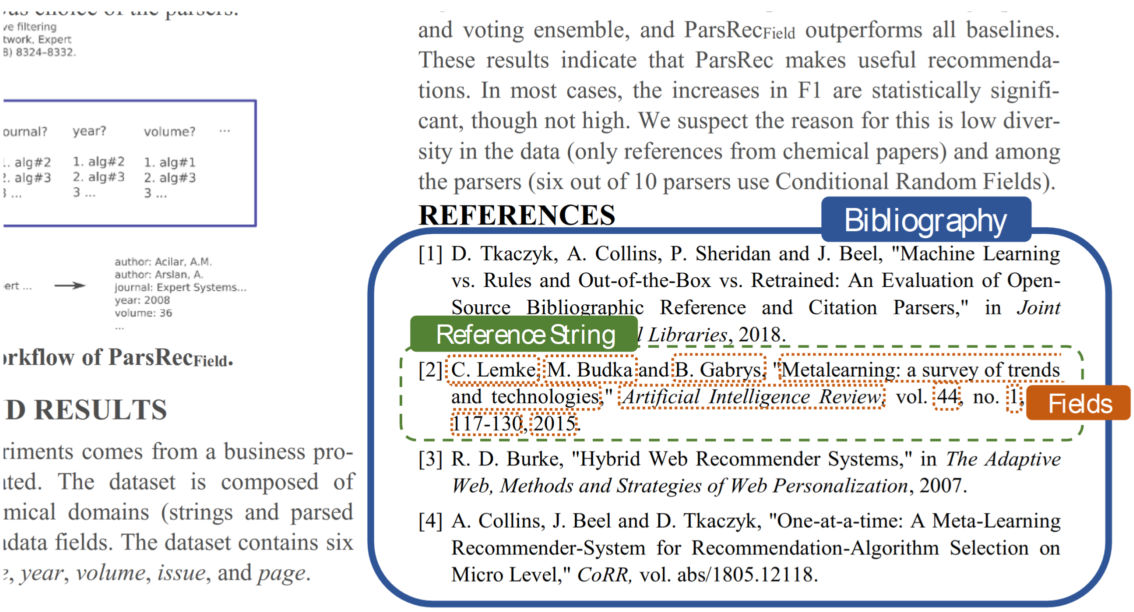

Figure 1: Illustration of a 'Bibliography' with four 'Reference Strings', each with a number of 'Fields'. A reference parser receives a reference string as input, and outputs labelled fields, e.g. authors="C. Lemke ..."; title="Metalearning: a survey ..."; ...

Harvard, ACM). The dataset was synthetically created. This means, the reference strings are not 'real' reference strings extracted from 'real' bibliographies. Instead, we downloaded 677,000 references in XML format from CrossRef, and used Citeproc-JS (Frank G. Bennett, 2011) with 1,500 citation styles to convert the 677,000 references into a total of 1 billion annotated citation strings (1,500 * 677,000).[2]

We wonder how suitable a synthetic dataset like GIANT is to train machine learning models for citation parsing. Therefore, we pursue the following research question:

1. How will citation parsing perform when trained on synthetic reference strings, compared to being trained on real reference strings?

Potentially, synthetic data could lead to higher citation parsing performance, as synthetic datasets may contain more data and more diverse data (more citation styles). Synthetic datasets like GIANT could potentially also advance (deep) citation parsing, which currently suffers from a lack of 'real' annotated bibliographies at large scale.

In addition to the above research question, we aimed to answer the following questions:

[2] We use the terms 'citation parsing', 'reference parsing', and 'reference-string parsing' interchangeably.

2. To what extent does citation-parsing (based on machine learning) depend on the amount of training data?

3. How important is re-training a citation parser for the specific data it should be used on? Or, in other words, how does performance vary if the test data differs (not) from the training data?

4. Is it important to have many different fields (author, year, ...) for training, even if the fields are not available in the final data?

2 Related Work

We are aware of eleven datasets (Figure 2) with annotated reference strings. The most popular ones are probably Cora and CiteSeer. Researchers also often use variations of PubMed. Several datasets are from the same authors, and many datasets include data from other datasets. For instance, the Grobid dataset is based on some data from Cora, PubMed, and others (Lopez, 2020). New data is continuously added to Grobid's dataset. As such, there is not "the one" Grobid dataset. GIANT (Grennan et al., 2019) is the largest and most diverse dataset in terms of citation styles, but GIANT is, as mentioned, synthetically created.

Cora is one of the most widely used datasets but has potential shortcomings (Anzaroot and Mc-

Callum, 2013; Councill et al., 2008; Prasad et al., 2018). Cora is homogeneous with citation strings only from Computer Science. It is relatively small and only has labels for "coarse-grained fields" (Anzaroot and McCallum, 2013). For example, the author field does not label each author separately. Prasad et al. conclude that a "shortcoming of [citation parsing research] is that the evaluations have been largely limited to the Cora dataset, which is [...] unrepresentative of the multilingual, multidisciplinary scholastic reality" (Prasad et al., 2018).

3 Methodology

To compare the effectiveness of synthetic vs. real bibliographies, we used Grobid. Grobid is the most effective citation parsing tool (Tkaczyk et al., 2018a) and the most easy to use tool based on our experience. Grobid uses conditional random fields (CRF) as machine learning algorithm. Of course, in the long-run, it would be good to conduct our experiments with different machine learning algorithms, particularly deep learning algorithms, but for now we concentrate on one tool and algorithm. Given that all major citation-parsing tools – including Grobid, Cermine and ParsCit – use CRF we consider this sufficient for an initial experiment. Also, we attempted to re-train Neural ParsCit (Prasad et al., 2018) but failed doing so, which indicates that the ease-of-use of the rather new deep-learning methods is not yet as advanced as the established citation parsing tools like Grobid.

We trained Grobid, the CRF respectively, on two datasets. $\text{Train}_{\text{Grobid}}$ denotes a model trained on 70% (5,460 instances) of the dataset that Grobid uses to train its out-of-the box version. We slightly modified the dataset, i.e. we removed labels for 'pubPlace', 'note' and 'institution' as this information is not contained in GIANT, and hence a model trained on GIANT could not identify these labels[3]. $\text{Train}_{\text{GIANT}}$ denotes the model trained on a random sample (5,460 instances) of GIANT's 991,411,100 labeled reference strings. Our expectation was that both models would perform similar, or, ideally, $\text{Train}_{\text{GIANT}}$ would even outperform $\text{Train}_{\text{Grobid}}$.

To analyze how the amount of training data affects performance, we additionally trained $\text{Train}_{\text{GIANT}}$, on 1k, 3k, 5k, 10k, 20k, and 40k instances of GIANT.

We evaluated all models on four datasets. $\text{Eval}_{\text{Grobid}}$ comprises of the remaining 30% of Grobid's dataset (2,340 reference strings). $\text{Eval}_{\text{Cora}}$ denotes the Cora dataset, which comprises, after some cleaning, of 1,148 labelled reference strings from the computer science domain. $\text{Eval}_{\text{GIANT}}$ comprises of 5,000 random reference strings from GIANT.

These three evaluation datasets are potentially not ideal as evaluations are likely biased towards one of the two trained models. Evaluating the models on $\text{Eval}_{\text{GIANT}}$ likely favors $\text{Train}_{\text{GIANT}}$ since the data for both $\text{Train}_{\text{GIANT}}$ and $\text{Eval}_{\text{GIANT}}$ is highly similar, i.e. it originates from the same dataset. Similarly, evaluating the models on $\text{Eval}_{\text{Grobid}}$ likely favors $\text{Train}_{\text{Grobid}}$ as $\text{Train}_{\text{Grobid}}$ was trained on 70% of the original Grobid dataset and this 70% of the data is highly similar to the remaining 30% that we used for the evaluation. Also, the Cora dataset is somewhat biased, because Grobid's dataset contains parts of Cora. We therefore created another evaluation dataset.

$\text{Eval}_{\text{WebPDF}}$ is our 'unbiased' dataset with 300 manually annotated citation strings from PDFs that we collected from the Web. To create $\text{Eval}_{\text{WebPDF}}$, we chose twenty different words from the homepages of some universities[4]. Then, we used each of the twenty words as a search term in Google Scholar. From each of these searches, we downloaded the first four available PDFs. Of each PDF, we randomly chose four citation strings. This gave approximately sixteen citation strings for each of the twenty keywords. In total, we obtained 300 citation strings. We consider this dataset to be a realistic, though relatively small, dataset for citation parsing in the context of a web-based academic search engine or recommender system.

We measure performance of all models with precision, recall, F1 (Micro Average) and F1 (Macro Average) on both field level and token level. We only report 'F1 Macro Average on field level' as all metrics led to similar results.

All source code, data (including the WebPDF dataset), images, and an Excel sheet with all results (including precision and recall and token level results) is available on GitHub

[3] This is a shortcoming of GIANT. However, the purpose of our current work is to generally compare 'real' vs synthetic data. Hence, both datasets should be as similar as possible in terms of available fields to make a fair comparison. Therefore, we removed all fields that were not present in both datasets.

[4] The words were: bone, recommender systems, running, war, crop, monetary, migration, imprisonment, hubble, obstetrics, photonics, carbon, cellulose, evolutionary, revolutionary, paleobiology, penal, leadership, soil, musicology.

Dataset Name	# Instances	Domain
Cora [29]	1,295	Computer Science
CiteSeer [16]	1,563	Artificial Intelligence
Umass [2]	1,829	STEM
FLUX-CiM CS [20]	300	Computer Science
FLUX-CiM HS [20]	2,000	Health Science
GROBID [26–28]	6,835	Multi-Domain (Cora, arXiv, PubMed...)
PubMed (Central) [9, 17]	Varies	Biomedical
GROTOAP2 (Cermine) [35–37]	6,858	Biomedical & Computer Science
CS-SW [20]	578	Semantic Web Conferences
Venice [33]	40,000	Humanities
GIANT [19]	991 million	Multi-Domain (~1,500 Citation Styles)

Figure 2: List of Citation Datasets

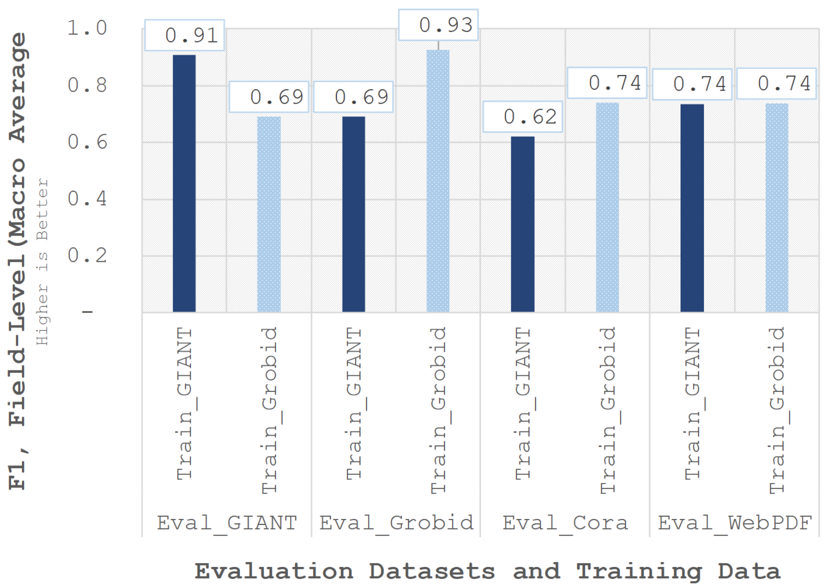

Figure 3: F1 of the two models ($Train_{Grobid}$ and $Train_{GIANT}$) on the four evaluation datasets.

https://github.com/BeelGroup/GIANT-The-1-Billion-Annotated-Synthetic-Bibliographic-Reference-String-Dataset/.

4 Results

The models trained on Grobid ($Train_{Grobid}$) and GIANT ($Train_{GIANT}$) perform as expected when evaluated on the three 'biased' datasets $Eval_{Grobid}$, $Eval_{Cora}$ and $Eval_{GIANT}$ (Figure 3). When evaluated on $Eval_{Grobid}$, $Train_{Grobid}$ outperforms $Train_{GIANT}$ by 35% with an F1 of 0.93 vs. 0.69. When evaluated on $Eval_{GIANT}$, results are almost exactly the opposite: This time, $Train_{GIANT}$ outperforms $Train_{Grobid}$ by 32% with an F1 of 0.91 vs. 0.69. On $Eval_{Cora}$, the difference is less strong but still notable. $Train_{Grobid}$ outperforms $Train_{GIANT}$ by 19% with an F1 of 0.74 vs. 0.62. This is not surprising as Grobid's training data includes some Cora data.

While these results generally might not be surprising, they imply that both synthetic and real data lead to very similar results and 'behave' similarly when used to train models that are evaluated on data being (not) similar to the training data.

Also interesting is the evaluation on the WebPDF dataset. The model trained on synthetic data ($Train_{GIANT}$) and the model trained on real data ($Train_{Grobid}$) perform alike with an F1 of 0.74 each (Figure 3)[5]. In other words, synthetic and human-labelled data perform equally well for training our machine learning models.

Looking at the data in more detail reveals that some fields are easier to parse than others (Figure 4). For instance, the 'date' field (i.e. year of publication) has a constantly high F1 across all models and evaluation datasets (min=0.86; max=1.0). The 'author' field also has a high F1 throughout all experiments (min=0.75; max=0.99). In contrast, parsing 'booktitle' and 'publisher' seems to strongly benefit from training based on samples similar to the evaluation data. When evaluation and training data is highly similar (e.g. $Train_{GIANT}$–$Eval_{GIANT}$ or $Train_{Grobid}$–$Eval_{Grobid}$), F1 is relatively high (typically above 0.7). If the evaluation data is different (e.g. $Train_{GIANT}$– $Eval_{Grobid}$), F1 is low (0.15 and 0.16 for $Train_{Grobid}$ and $Train_{GIANT}$ respectively on $Eval_{WebPDF}$). The difference in F1 for parsing the book-title is around factor 6.5, with an F1 of 0.97 ($Train_{Grobid}$) and 0.15 respectively ($Train_{GIANT}$) when evaluated on $Eval_{Grobid}$.

Similarly, F1 for parsing the book-title on $Eval_{GIANT}$ differs by around factor 3 with an F1 of 0.75 ($Train_{GIANT}$) and 0.27 ($Train_{Grobid}$) respectively. While it is well known, and quite intuitive, that different fields are differently difficult to parse, we are first to show that field accuracy varies for different fields differently depending on whether or not the model was trained on data (not) being similar to the evaluation data.

In a side experiment, we trained a new model $Train_{Grobid}$+ with additional labels for institution, note and pubPlace (those we removed for the other experiments). $Train_{Grobid}$+ outperformed $Train_{Grobid}$ notably with an F1 of 0.84 vs. 0.74 (+13.5%) when evaluated on $Eval_{WebPDF}$. This indicates that the more fields are available for training, the better the parsing of all fields becomes even if the additional fields are not in the evaluation data. This finding seems plausible to us and confirms statements by Anzaroot and McCallum but, to the best of our knowledge, we are first to quantify the benefit. It is worth noting that citation parsers do not always use the same fields (Figure 6). For instance, Cermine extracts relatively few fields, but is one of few tools extracting the DOI field.

Our assumption that more training data would generally lead to better parsing performance – and hence GIANT could be useful for training standard machine learning algorithms – was not confirmed. Increasing training data from 1,000 to 10,000 instances improved F1 by 6% on average over the four evaluation datasets (Figure 5). More precisely, increasing data from 1,000 to 3,000 instances improved F1, on average, by 2.4%; Increasing from 3,000 to 5,000 instances improved F1 by another 2%; Increasing further to 10,000 instances improved F1 by another 1.6%. However, increasing to 20,000 or 40,000 instances leads to no notable improvement, and in some cases even to a decline in F1 (Figure 5).

5 Summary and Discussion

In summary, both models – one trained on synthetic data (GIANT) and one trained on 'real' human-annotated reference strings (Grobid) – performed very similar. On the main evaluation dataset (WebPDF) both models achieved an F1 of 0.74. Similarly, if a model was evaluated on data different from its training data, F1 was between 0.6 and

[5] All results are based on the Macro Average F1. Looking at the Micro Average F1 shows a slightly better performance for $Train_{Grobid}$ than for $Train_{GIANT}$ (0.82 vs. 0.80), but the difference is neither large nor statistically significant (p¡0.05).

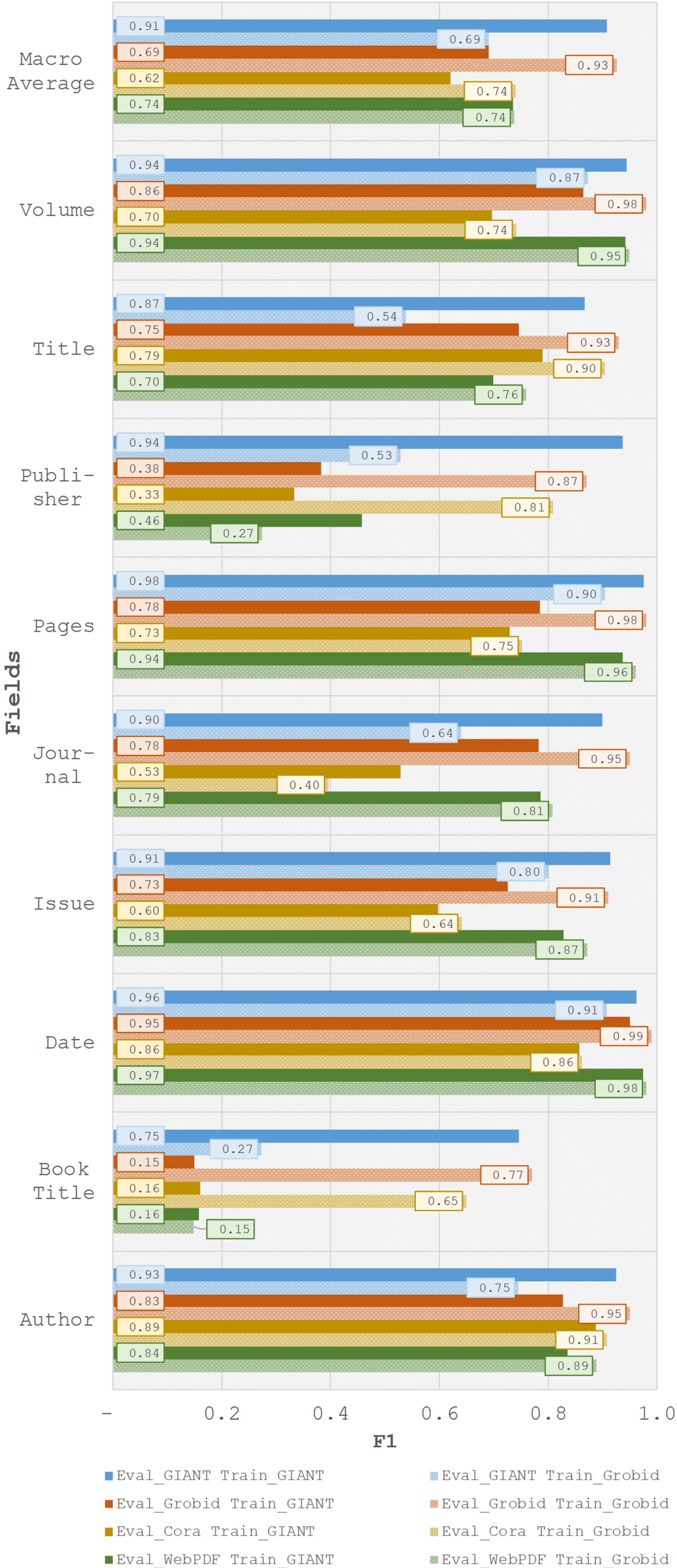

Figure 4: F1 for different fields (title, author, ...), evaluation dataset and training data.

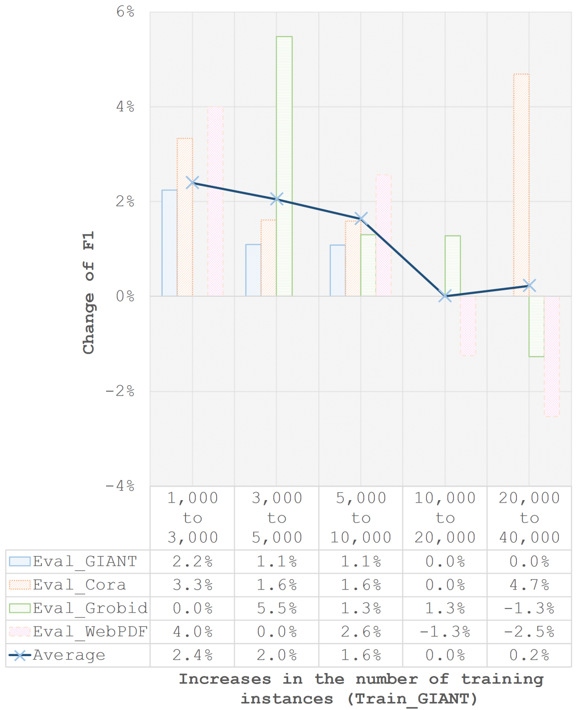

Figure 5: Performance (F1) of $\text{Train}_{\text{GIANT}}$ on the four evaluation datasets, by the number of training instances.

Citation Parser	Approach	Extracted Fields
Biblio	Regular Expressions	author, date, editor, genre, issue, pages, publisher, title, volume, year
BibPro	Template Matching	author, title, venue, volume, issue, page, date, journal, booktitle, techReport
CERMINE	Machine Learning (CRF)	author, issue, pages, title, volume, year, DOI, ISSN
GROBID		authors, booktitle, date, editor, issue, journal, location, note, pages, publisher, title, volume, web, institution
ParsCit		author, booktitle, date, editor, institution, journal, location, note, pages, publisher, tech, title, volume
Neural ParsCit	Deep Learning	author, booktitle, date, editor, institution, journal, location, note, pages, publisher, tech, title, volume

Figure 6: The approach and extracted fields of six popular open-source citation parsing tools

0.7. If a model was evaluated on data similar to the training data, F1 was above 0.9 (+30%). F1 only increased up to a training size of around 10,000 instances (+6% compared to 1,000 instances). Additional fields (e.g. pubplace) in the training data increased F1 notably (+13.5%), even if these additional fields were not in the evaluation data.

These results lead us to the following conclusions.

First, there seems to be little benefit in using synthetic data (e.g. GIANT (Grennan et al., 2019)) for training traditional machine learning models (i.e. conditional random fields). The existing datasets with a few thousand training instances seem sufficient.

Second, citation parsers should, if possible, be (re)trained on data that is similar to the data that should actually be parsed. Such a re-training increased performance by around 30% in our experiments. This finding may also explain why researchers often report excellent performance of their tools and approaches with e.g. F1's of over 0.9. These researchers typically evaluate their models on data highly similar to the training data. This might be considered a realistic scenario for those cases when re-training is possible. However, reporting such results creates unrealistic expectations for scenarios without the option to re-train, i.e. for users who just want to use a citation parser like Grobid out-of-the-box. Therefore, we propose that future evaluations of citation parsing algorithms should be conducted on at least two datasets: One dataset that is similar to the training dataset, and one out-of-sample dataset that differs from the training data.

Third, citation parsers should be trained with as many labelled field types as possible, even if these fields will not be in the data that should be parsed. Such a fine-grained training improved F1 by 13.5% in our experiments.

Fourth, having ten times as much training data (10,000 vs. 1,000) improved the parsing performance by 6%, without notable improvements beyond 10,000 instances. Annotating a few thousand instances should be feasible for many scenarios. Hence, businesses and organizations who want the maximum accuracy should annotate their own data for training as this likely will lead to large increases in accuracy (+30%, see conclusion 3).

Fifth, given how similar synthetic and traditionally annotated data perform, synthetic data likely is suitable to train deep neural networks for citation parsing. This, of course, has yet to be empirically to be shown. However, if our assumption holds true, deep citation parsers could greatly benefit from synthetic data like GIANT.

For the future, we see the need to extend our experiments to different machine learning algorithms and datasets (e.g. unarXive (Saier and Färber, 2020) or CORE (Knoth and Zdrahal, 2012)). It would also be interesting to analyze if and to what extend synthetic data could improve related disciplines. This may include citation-string matching, i.e. analyzing whether two different reference strings refer to the same document (Ghavimi et al., 2019), or the extraction of mathematical formulae (Greiner-Petter et al., 2020) or titles (Lipinski et al., 2013) from scientific articles.

Acknowledgments

We are grateful for the support received by Martin Schibel, Andrew Collins and Dominika Tkaczyk in creating the GIANT dataset. We would also like to acknowledge that this research was partly conducted with the financial support of the ADAPT SFI Research Centre at Trinity College Dublin. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106.

References

Dong An, Liangcai Gao, Zhuoren Jiang, Runtao Liu, and Zhi Tang. 2017. Citation metadata extraction via deep neural network-based segment sequence labeling. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1967–1970. ACM.

Sam Anzaroot and Andrew McCallum. 2013. A new dataset for fine-grained citation field extraction. *ICML Workshop on Peer Reviewing and Publishing Models*.

Nisa Bakkalbasi, Kathleen Bauer, Janis Glover, and Lei Wang. 2006. Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries*, 3.

Joeran Beel and Bela Gipp. 2009a. Google Scholar's Ranking Algorithm: An Introductory Overview. In *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, volume 1, pages 230–241, Rio de Janeiro (Brazil). International Society for Scientometrics and Informetrics. Available at http://docear.org.

Joeran Beel and Bela Gipp. 2009b. Google Scholar's Ranking Algorithm: The Impact of Citation Counts (An Empirical Study). In *Proceedings of the 3rd IEEE International Conference on Research Challenges in Information Science (RCIS'09)*, pages 439–446, Fez (Morocco). IEEE. Available at http://docear.org.

Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. 2016. Research paper recommender systems: A literature survey. *International Journal on Digital Libraries*, (4):305–338.

Akansha Bhardwaj, Dominik Mercier, Andreas Dengel, and Sheraz Ahmed. 2017. Deepbibx: Deep learning for image based bibliographic data extraction. In *International Conference on Neural Information Processing*, pages 286–293. Springer.

Guillaume Cabanac, Ingo Frommholz, and Philipp Mayr. 2020. Bibliometric-enhanced information retrieval (bir) 10th anniversary workshop edition. *arXiv preprint arXiv:2001.10336*.

I.G. Councill, C.L. Giles, and M.Y. Kan. 2008. ParsCit: An open-source CRF reference string parsing package. In *Proceedings of LREC*, volume 2008, pages 661–667. European Language Resources Association (ELRA).

Masaki Eto. 2019. Extended co-citation search: Graph-based document retrieval on a co-citation network containing citation context information. *Information Processing & Management*, 56(6):102046.

Michael Färber and Adam Jatowt. 2020. Citation recommendation: Approaches and datasets. *arXiv preprint arXiv:2002.06961*.

Michael Färber, Alexander Thiemann, and Adam Jatowt. 2018. Citewerts: A system combining cite-worthiness with citation recommendation. In *European Conference on Information Retrieval*, pages 815–819. Springer.

Jr. Frank G. Bennett. 2011. The citeproc-js Citation Processor.

Behnam Ghavimi, Wolfgang Otto, and Philipp Mayr. 2019. An evaluation of the effect of reference strings and segmentation on citation matching. In *International Conference on Theory and Practice of Digital Libraries*, pages 365–369. Springer.

André Greiner-Petter, Moritz Schubotz, Fabian Müller, Corinna Breitinger, Howard S Cohl, Akiko Aizawa, and Bela Gipp. 2020. Discovering mathematical objects of interest–a study of mathematical notations. *arXiv preprint arXiv:2002.02712*.

Mark Grennan, Martin Schibel, Andrew Collins, and Joeran Beel. 2019. Giant: The 1-billion annotated synthetic bibliographic-reference-string dataset for deep citation parsing. In *27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science*, pages 101–112.

P. Jacso. 2008. Testing the calculation of a realistic h-index in Google Scholar, Scopus, and Web of Science for FW Lancaster. *Library Trends*, 56(4):784–815.

Haofeng Jia and Erik Saule. 2018. Graph embedding for citation recommendation. *arXiv preprint arXiv:1812.03835*.

Petr Knoth and Zdenek Zdrahal. 2012. Core: three access levels to underpin open access. *D-Lib Magazine*, 18(11/12).

Mario Lipinski, Kevin Yao, Corinna Breitinger, Joeran Beel, and Bela Gipp. 2013. Evaluation of header metadata extraction approaches and tools for scientific pdf documents. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries (JCDL'13)*, pages 385–386.

Avishay Livne, Vivek Gokuladas, Jaime Teevan, Susan T Dumais, and Eytan Adar. 2014. Citesight: supporting contextual citation recommendation using differential search. pages 807–816.

Patrice Lopez. 2009. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *International conference on theory and practice of digital libraries*, pages 473–474. Springer.

Patrice Lopez. 2013. Grobid, github repository. *https://github.com/kermitt2/grobid/*.

Patrice Lopez. 2020. Training data query #535. *GitHub https://github.com/kermitt2/grobid/issues/535*.

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2018. Information extraction from scientific articles: a survey. *Scientometrics*, 117(3):1931–1990.

Animesh Prasad, Manpreet Kaur, and Min-Yen Kan. 2018. Neural parscit: a deep learning-based reference string parser. *International Journal on Digital Libraries*, 19(4):323–337.

Syed Tahseen Raza Rizvi, Andreas Dengel, and Sheraz Ahmed. 2019. Deepbird: An automatic bibliographic reference detection approach. *arXiv preprint arXiv:1912.07266*.

Danny Rodrigues Alves, Giovanni Colavizza, and Frédéric Kaplan. 2018. Deep reference mining from scholarly literature in the arts and humanities. *Frontiers in Research Metrics and Analytics*, 3:21.

Tarek Saier and Michael Färber. 2020. unarxive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata. *Scientometrics*, pages 1–24.

Dominika Tkaczyk, Andrew Collins, Paraic Sheridan, and Joeran Beel. 2018a. Machine learning vs. rules and out-of-the-box vs. retrained: An evaluation of open-source bibliographic reference and citation parsers. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, pages 99–108, New York, NY, USA. ACM.

Dominika Tkaczyk, Rohit Gupta, Riccardo Cinti, and Joeran Beel. 2018b. Parsrec: A novel meta-learning approach to recommending bibliographic reference parsers. In *Proceedings of the 26th Irish Conference on Artificial Intelligence and Cognitive Science (AICS)*, volume 2259, pages 162–173. CEUR-WS.

Dominika Tkaczyk, Paraic Sheridan, and Joeran Beel. 2018c. Parsrec: A meta-learning recommender system for bibliographic reference parsing tools. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys)*.

Dominika Tkaczyk, Paweł Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, and Łukasz Bolikowski. 2015. Cermine: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(4):317–335.

Yiqing Zhang. 2018. Towards highly accurate publication information extraction from academic homepages.