Bootstrapping Multilingual Metadata Extraction: A Showcase in Cyrillic

Johan Krause and **Igor Shapiro** and **Tarek Saier** and **Michael Färber**
Institute AIFB, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
{johan.krause,igor.shapiro}@student.kit.edu
{tarek.saier,michael.faerber}@kit.edu

Abstract

Applications based on scholarly data are of ever increasing importance. This results in disadvantages for areas where high-quality data and compatible systems are not available, such as non-English publications. To advance the mitigation of this imbalance, we use Cyrillic script publications from the CORE collection to create a high-quality data set for metadata extraction. We utilize our data for training and evaluating sequence labeling models to extract title and author information. Retraining GROBID on our data, we observe significant improvements in terms of precision and recall and achieve even better results with a self developed model. We make our data set covering over 15,000 publications as well as our source code freely available.[1]

1 Introduction

The use of scholarly data becomes more and more important as the rate of academic publications keeps increasing and automated processing gains relevance, such as scientometric analysis and scholarly recommendation (Sigurdsson, 2020; Zhang et al., 2020). Consequentially, limitations of scholarly data and approaches based thereon directly translate into disadvantages for the affected publications, in terms of, for example, discoverability and impact. One particular limitation of scholarly data nowadays is an underrepresentation of non-English content (Vera-Baceta et al., 2019; Moskaleva and Akoev, 2019). While supporting multiple languages poses challenges, such as language-specific preprocessing requirements (Grave et al., 2018; McCann, 2020), disregarding non-English work is problematic (Amano et al., 2016; Lynch et al., 2021). To further the availability of high-quality scholarly data beyond the anglophone publication record, we showcase the creation and application of a data set for training and evaluating sequence labeling tasks on Cyrillic publications.

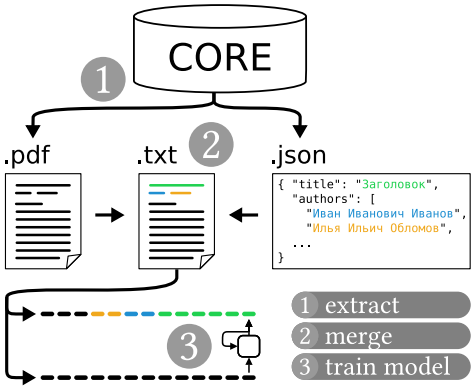


Figure 1: Schematic overview of our approach.

Recent years have seen an increased focus on multilinguality in natural language processing approaches, such as language models (Devlin et al., 2019) and data sets (Caswell et al., 2021). Furthermore, there are efforts to specifically support languages that use non-Latin scripts (Roark et al., 2020; Pfeiffer et al., 2021). With regards to Cyrillic script languages, approaches concerned with named entity linking in Web documents (Piskorski et al., 2021), as well as approaches to extracting keywords from scientific texts (Bolshakova et al., 2019) exist. Model training for these types of information extraction tasks is increasingly done using automatically generated high-quality training data. This has, for example, been done for tasks such as text extraction from scholarly PDF files (Bast and Korzen, 2017), identification of publication components such as figures and tables in scanned documents (Ling and Chen, 2020), and the parsing of bibliographic references (Grennan and Beel, 2020; Thai et al., 2020).

We extend this approach to non-English scholarly data. To this end, we use Cyrillic script documents from the CORE data set (Knoth and Zdrahal, 2012) to train and evaluate sequence labeling mod-

[1] See https://github.com/IllDepence/sdp2021.

Proceedings of the Second Workshop on Scholarly Document Processing, pages 66–72
June 10, 2021. ©2021 Association for Computational Linguistics

els for identifying publications' metadata (title and authors) in unlabeled text, as illustrated in Figure 1.

Overall, the contributions we make with this paper are as follows.

1. We showcase an effective method for creating high-quality data for training and evaluating metadata extraction sequence labeling models on multilingual scholarly data.

2. We provide a data set for Cyrillic, comprising 15,553 publications spanning three languages and 27 years.

3. We create sequence labeling models that outperform available methods on Cyrillic data.

2 Data Set Creation

2.1 Data Selection

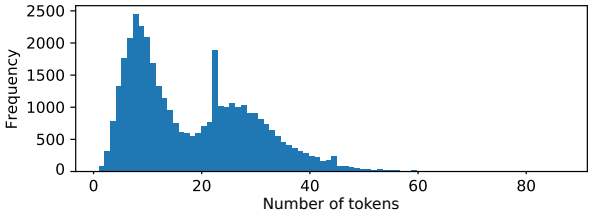
Although many large scholarly data sets exist nowadays, most are restricted in terms of language coverage, language related metadata, or availability of full text documents. The PubMed Central Open Access Subset,[2] for example, only contains Latin script publications,[3] the Semantic Scholar Open Research Corpus (Lo et al., 2020) is restricted to English, and the Microsoft Academic Graph (Sinha et al., 2015; Wang et al., 2019) contains no full texts. Furthermore, none of the aforementioned offers metadata on publications' language. We chose to use the CORE data set[4] (Knoth and Zdrahal, 2012)—a large scholarly data set consisting of PDF documents and metadata aggregated from institutional and subject repositories—for our approach because it is not restricted by language, offers full papers and partly provides language metadata.

To obtain Cyrillic script publications, we first filter the whole collection for the language labels of four Cyrillic script languages, namely Russian, Ukrainian, Bulgarian, and Macedonian, resulting in 23,850 documents. Noticing that a lot of the items we identified are clustered in certain ID ranges of CORE, we extend our data to roughly 48,000 papers by applying language detection on the PDF files of documents adjacent in the set of CORE IDs. After removal of duplicates (papers with different CORE ID but identical PDF) we end up with 27,755 documents.

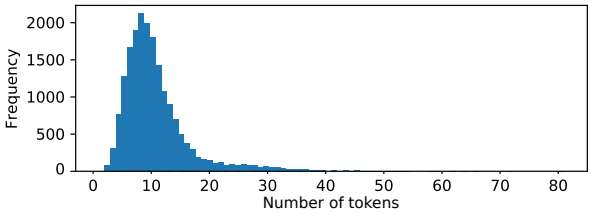
[2]See https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/.

[3]See https://www.ncbi.nlm.nih.gov/pmc/about/faq/#q16.

[4]Specifically, we use the *2018-03-01 full text data set* version of CORE containing 123,988,821 documents.



(a) Distribution before keyword filtering.



(b) Distribution after keyword filtering.

Figure 2: Change in document title length due to keyword filtering.

Examination of our data at this point reveals that it contains documents other than scientific papers, such as lecture notes, lecture schedules, and untypically long documents such as whole conference proceedings. To remove these, we perform two filtering steps. First, we remove documents whose title contains either of the words студентів (UKR: "student"), Конспект лекцій (UKR: "lecture schedule"), Програма (RUS: "program", as in study program) and Диплом (RUS: "diploma"), leaving around 22,000 documents and changing the distribution of document title lengths as shown in Figure 2. Second, we drop documents whose length exceeds the 95% quantile (68 pages). Finally, we remove papers for which CORE does not provide basic metadata, and papers for which the plain text was not extractable from the PDF. This leaves us with 15,553 papers, which form the basis for our work and the provided Cyrillic data set.

2.2 Data Preparation

To prevent having to remove large portions of the identified Cyrillic papers due to missing metadata (see previous section), we decide to focus on publications' *title* and list of *authors*. In order to create training data for sequence labeling tasks, we obtain the JSON metadata and PDF of each of the selected publications from CORE. From the PDF, we extract the plain text contained in the first page using *PDFMiner*[5], identify the title and authors from the JSON metadata and insert labels accordingly (see Section 3.2.1 for details).

[5]See https://github.com/euske/pdfminer.

2.3 Data Set

The resulting data set comprises *15,553 papers* spanning *27 years* and *three languages*. For each paper, we provide ground truth sequence labeling output in TEI[6] format and as annotated plain text.[7]

A detailed breakdown of languages, obtained using fastText (Joulin et al., 2016, 2017) language detection is shown in Table 1. Languages with less than five occurrences throughout the data set are not included. The distribution of papers by publication year is shown in Figure 3. A breakdown of the topics[8] covered by the data set is shown in Table 2. Analysing the origin of papers, we note that 90% originate from either the "A.N.Beketov KNUME Digital Repository"[9] or the "Zhytomyr State University Library."[10]

Language	#Documents
Ukrainian	11,708
Russian	3,786
Bulgarian	54

Table 1: Distribution of languages.

Topic	#Documents
Engineering	2,472
Economics	2,429
Urban Planning/Infrastructure	2,263
Education	2,255
Other (Linguistics, Zoology, Psychology ...)	6,134

Table 2: Distribution of topics.

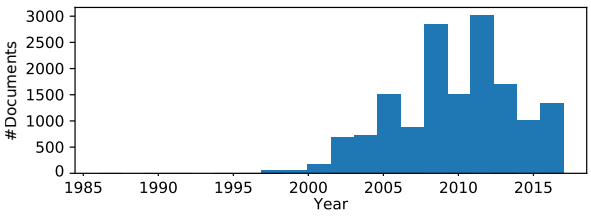


Figure 3: Distribution of publication years of the final data set.

[6]See https://tei-c.org/.

[7]See https://zenodo.org/record/4708696.

[8]For details of how topics were determined see https://github.com/IllDepence/sdp2021.

[9]See https://eprints.kname.edu.ua/.

[10]See http://eprints.zu.edu.ua/.

3 Application

To assess the utility of our data set, we use it to retrain GROBID (Lopez, 2008–2021), a widely used metadata extraction tool (Nasar et al., 2018), as well as a standalone sequence labeling model, and evaluate their performance against an off-the-shelf version of GROBID.

3.1 GROBID Training

GROBID utilizes several models for different tasks, each of which can be retrained. Our use case—the extraction of title and author information—concerns the *header* model, which is based on conditional random fields (CRF). Retraining the header model from scratch using our data set, we note that for a significant portion of PDFs, GROBID is not able to produce plain text on which the CRF would then be applied. Because of this, we are only able to use 9,620 papers (62% of the data set) for retraining.

3.2 Standalone Sequence Labeling Model

3.2.1 Data Preprocessing

For our standalone model we decide to label the textual content of the first page of each paper using four tags, namely *Author*, *B-title* (beginning of the title, i.e. the first title token), *I-title* (tokens inside the title) and *Misc* (everything else).

To this end, we extract the plain text from the PDF using *PDFMiner*, tokenize the text according to whitespace, and replace newlines with a `NEWLINE` token. The publication's title is then identified using the JSON metadata and each token labeled accordingly. `NEWLINE` tokens within a sequence of title tokens are preserved.

For the matching of authors, we split the author strings from the metadata into surname and given names. We first locate the surnames in the token sequence, and label the occurrence closest to the title as *Author*. Because given names can appear written-out as well as abbreviated in the form of initials, we heuristically identify the latter as follows. Given an identified surname, we search within a window of eight tokens before and after the surname[11] for uppercase characters followed by a period. Matching initials are then labeled accordingly. Written-out given names are normally

[11]Eight being given in the edge case where a surname is followed by a separating comma, two initials and a newline somewhere in-between. E.g.: "<surname>,␣<initial>.␣<newline><initial>.".

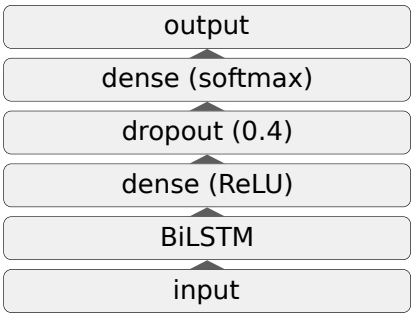


Figure 4: Network architecture.

matched just like surnames.

From the tokens we derive vectorized embeddings using *fastText*. Following Chiu and Nichols (2016) we use representations with 100 dimensions. In addition to the embeddings, we add five additional feature dimensions to the word vectors as done by Huang et al. (2015). These contain information about whether a token is uppercase, capitalized, contains punctuation, contains a line break or is styled like an author initial (uppercase and ending in a period character).

3.2.2 Model Training

For our standalone model we choose to use a BiLSTM network, as is commonly done for sequence labeling tasks (Huang et al., 2015).

We trim input sequences to the first 1,000 tokens, resulting in an input space of $1,000 \times 105$ dimensions per document, as each token is represented by a 100-dimensional vector with a set of five added features per token. The output space is of equal length and contains a one-hot-encoded representation of one of the four labels *Author*, *B-title*, *I-title* and *Misc*.

Because title and authors only make up a small fraction of the words at the beginning of a publication, tokens with the *Misc* label make up a majority of our data. To prevent the trivial prediction of the *Misc* label playing too much of a role in training, each input word token is given an individual, heuristically determined weight value of either 1 for *Misc.* or 5 for *Author* and **-title* labels.

The final network, as shown in Figure 4, consists of a BiLSTM layer followed by a ReLU activated dense layer, a dropout layer and a final dense layer with softmax activation. For training, categorical cross entropy serves as the model's loss function and recall is employed as the target metric. Furthermore, the Adam optimizer (Kingma and Ba, 2017) with a learning rate of 0.0001 is used.

Model	Precision	Recall	F1
GROBID vanilla	0.06	0.06	0.06
GROBID retrained	**0.85**	0.81	0.83
BiLSTM	0.84	**0.96**	**0.90**

Table 3: Overall evaluation scores.

$\text{Model}_{\text{label}}$	Precision	Recall	F1
GROBID $\text{retr.}_{\text{title}}$	**0.90**	0.90	0.90
$\text{BiLSTM}_{\text{title}}$	0.88	**0.96**	**0.92**
GROBID $\text{retr.}_{\text{author}}$	**0.81**	0.74	0.77
$\text{BiLSTM}_{\text{author}}$	0.80	**0.95**	**0.87**
GROBID $\text{retr.}_{\text{misc}}$	-	-	-
$\text{BiLSTM}_{\text{misc}}$	0.99	0.99	0.99

Table 4: Evaluation scores per label.

4 Evaluation

To assess the performance of both the off-the-shelf and retrained GROBID as well as the standalone BiLSTM model, we perform five-fold cross-validations and measure the overall precision, recall, and F1 score.[12]

Because GROBID retraining is only possible on roughly two thirds of our data (see Section 3.1) we evaluate the off-the-shelf ("vanilla") GROBID model on the same subset in order to maximize comparability of the evaluation results.

Regarding the comparability to our standalone BiLSTM model, a key difference lies in the fact that we use four labels (*Author*, *B-title*, *I-title* and *Misc*) instead of GROBID's two (*Author* and *Title*). To adjust for this difference, we decide to disregard the *Misc* label and combine the two types of **-title* label by a weighted average.

The overall evaluation scores resulting from this are shown in Table 3. We note that off-the-shelf GROBID is only able to determine a small fraction of title and author tokens correctly. Retraining GROBID using our training data, however, significantly improves the performance from an F1 score of 0.06 to 0.83, on par with GROBID's performance on English documents (Nasar et al., 2018). Our standalone BiLSTM model outperforms the retrained GROBID due to significantly higher recall with a F1 score of 0.90. Looking at the evaluation results per label for the retrained GROBID and standalone BiLSTM model, as shown in Table 4, we can see that the largest performance difference

[12] Since off-the-shelf GROBID does not have to be retrained, it is simply evaluated on 100% of the data instead of five folds.

Language	Precision	Recall	F1
Ukrainian	0.83	0.95	0.89
Russian	**0.88**	**0.97**	**0.92**
Bulgarian	0.51	0.70	0.58

Table 5: BiLSTM evaluation scores per language.

is given in the recall of the author label (measuring 0.74 and 0.95 respectively).

For further assessment of the BiLSTM model's performance, we evaluate its predictions per language as shown in Table 5. We can observe that the model achieves higher scores for Russian documents compared to the results for Ukrainian. This is especially notable since the amount of Ukrainian documents in the data set is significantly higher than that of Russian papers. One possible explanation of this performance gap could be a more consistent structure among the Russian documents. Performance on the 50 Bulgarian documents within the data set is comparatively low. While this could likely be due to the vast majority of the respective training data being in a different language, the informativeness of the score itself has to be considered keeping in mind that there are merely 50 documents for testing available.

5 Conclusion

Inspired by recent approaches creating high-quality data for training and evaluating information extraction tasks involving scholarly publications, we utilize this approach to tackle the problem of underrepresented non-English scholarly (training) data. To this end, we use Cyrillic script documents found in the CORE data set to train sequence labeling models for identifying publications' metadata.

We create a data set of *15,553 papers* spanning *27 years* and *three languages*. Using this data set, we retrain GROBID and thereby greatly improve its performance. Furthermore, we train and evaluate a separate sequence labeling model that is less constrained by PDF parsing restrictions (see Section 3.1), showing even better overall performance results than the retrained GROBID model.

By showcasing the use of freely available non-English publications to improve the availability of high-quality data and models covering areas beyond the anglophone publication record, we hope to inspire similar efforts for other languages. For our own approach, we plan to extend it to the extraction of bibliographic references in the future.

Author Contributions

Johan Krause and Igor Shapiro: Data curation, Formal Analysis, Investigation, Methodology, Software, Writing – original draft. Tarek Saier: Conceptualization, Data curation, Supervision, Writing – original draft, Writing – review & editing. Michael Färber: Supervision, Writing – review & editing.

References

Tatsuya Amano, Juan P. González-Varo, and William J. Sutherland. 2016. Languages are still a major barrier to global science. *PLOS Biology*, 14(12):1–8.

Hannah Bast and Claudius Korzen. 2017. A Benchmark and Evaluation for Text Extraction from PDF. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–10.

Elena Bolshakova, Natalia Efremova, and Kirill Ivanov. 2019. Terminological information extraction from russian scientific texts: Methods and applications. In *Proceedings of Third Workshop "Computational linguistics and language science"*, volume 4, pages 95–106.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Auguste Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara E. Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Javier Ortiz Suárez, Iroro Fred Ọ̀nọ̀mẹ̀ Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. In *Proceedings of the AfricaNLP Workshop*.

Jason P.C. Chiu and Eric Nichols. 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association*

for Computational Linguistics, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages.

Mark Grennan and Joeran Beel. 2020. 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. In *Proceedings of the 8th International Workshop on Mining Scientific Publications*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *Baidu research*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization.

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

Meng Ling and Jian Chen. 2020. DeepPaperComposer: A Simple Solution for Training Data Preparation for Parsing Research Papers. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 91–96. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983. Association for Computational Linguistics.

Patrice Lopez. 2008–2021. GROBID. `https://github.com/kermitt2/grobid`.

Abigail J. Lynch, Álvaro Fernández-Llamazares, Ignacio Palomo, Pedro Jaureguiberry, Tatsuya Amano, Zeenatul Basher, Michelle Lim, Tuyeni Heita Mwampamba, Aibek Samakov, and Odirilwe Selomane. 2021. Culturally diverse expert teams have yet to bring comprehensive linguistic diversity to intergovernmental ecosystem assessments. *One Earth*, 4(2):269–278.

Paul McCann. 2020. fugashi, a Tool for Tokenizing Japanese in Python. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 44–51, Online. Association for Computational Linguistics.

Olga Moskaleva and Mark Akoev. 2019. Non-English language publications in Citation Indexes - quantity and quality. In *Proceedings 17th International Conference on Scientometrics & Informetrics*, volume 1, pages 35–46, Italy. Edizioni Efesto.

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

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs Everywhere: Adapting Multilingual Language Models to New Scripts.

Jakub Piskorski, Bogdan Babych, Zara Kancheva, Olga Kanishcheva, Maria Lebedeva, Michał Marcińczuk, Preslav Nakov, Petya Osenova, Lidia Pivovarova, Senja Pollak, Pavel Přibáň, Ivaylo Radev, Marko Robnik-Sikonja, Vasyl Starko, Josef Steinberger, and Roman Yangarber. 2021. Slav-NER: the 3rd Cross-lingual Challenge on Recognition, Normalization, Classification, and Linking of Named Entities across Slavic Languages. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 122–133, Kiyv, Ukraine. Association for Computational Linguistics.

Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Isin Demirsahin, and Keith Hall. 2020. Processing South Asian Languages Written in the Latin Script: the Dakshina Dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2413–2423, Marseille, France. European Language Resources Association.

Steinn Sigurdsson. 2020. The future of arXiv and knowledge discovery in open science. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 7–9, Online. Association for Computational Linguistics.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. 2015. An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 243–246. ACM.

Dung Thai, Zhiyang Xu, Nicholas Monath, Boris Veytsman, and Andrew McCallum. 2020. Using bibtex to automatically generate labeled data for citation field extraction. In *Automated Knowledge Base Construction*.

Miguel-Angel Vera-Baceta, Michael Thelwall, and Kayvan Kousha. 2019. Web of Science and Scopus language coverage. *Scientometrics*, 121(3):1803–1813.

Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Darrin Eide, Yuxiao Dong, Junjie Qian, Anshul Kanakia, Alvin Chen, and Richard Rogahn. 2019. A Review of Microsoft Academic Services for Science of Science Studies. *Frontiers in Big Data*, 2:45.

Yu Zhang, Min Wang, Morteza Saberi, and Elizabeth Chang. 2020. Knowledge fusion through academic articles: a survey of definitions, techniques, applications and challenges. *Scientometrics*, 125(3):2637–2666.