Sequence Labeling for Citation Field Extraction from Cyrillic Script References

Igor Shapiro, Tarek Saier, Michael Färber

Institute AIFB, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
igor.shapiro@student.kit.edu, tarek.saier@kit.edu, michael.faerber@kit.edu

Abstract

Extracting structured data from bibliographic references is a crucial task for the creation of scholarly databases. While approaches, tools, and evaluation data sets for the task exist, there is a distinct lack of support for languages other than English and scripts other than the Latin alphabet. A significant portion of the scientific literature that is thereby excluded consists of publications written in Cyrillic script languages. To address this problem, we introduce a new multilingual and multidisciplinary data set of over 100,000 labeled reference strings. The data set covers multiple Cyrillic languages and contains over 700 manually labeled references, while the remaining are generated synthetically. With random samples of varying size of this data, we train multiple well performing sequence labeling BERT models and thus show the usability of our proposed data set. To this end, we showcase an implementation of a multilingual BERT model trained on the synthetic data and evaluated on the manually labeled references. Our model achieves an F1 score of 0.93 and thereby significantly outperforms a state-of-the-art model we retrain and evaluate on our data.

1 Introduction

Citations are a crucial part of the scientific discourse and represent a measure of the extent to which authors indirectly communicate with other researchers through publications (Shaw 1981). Therefore, accurate citation data is important for applications such as academic search engines (Ortega 2014) and academic recommender systems (e.g., for recommending papers (Beel et al. 2016) or citations (Färber and Jatowt 2020)). Since the number of scientific publications that is available on the web is growing exponentially (Khabsa and Giles 2014), it is crucial to automatically extract citation data from them. Many tools and models have been developed for this purpose, such as GROBID (Lopez 2009), CERMINE (Tkaczyk et al. 2015), and NEURAL PARSCIT (Prasad, Kaur, and Kan 2018). These tools mostly use supervised deep neural models. Accordingly, a large amount of labeled data is needed for training. However, most reference data sets are restricted in terms of discipline coverage and size, containing only several thousand instances (see Table 1). Furthermore, most models and tools are only trained on English data (Grennan et al. 2019; Prasad, Kaur, and Kan 2018). Therefore, existing models perform insufficiently on data in languages other than English, especially in languages written in scripts other than the Latin alphabet.

While English is the language with the largest share of scholarly literature, with estimates of over one hundred million documents (Khabsa and Giles 2014), other languages still make up a significant portion. For Russian alone, for example, there exist over 25 million scholarly publications (Moskaleva et al. 2018). Publications written in Cyrillic script languages, accordingly, make up an even larger portion, as they include further languages such as Ukrainian and Belarusian. A lack of methods and tools able to automatically extract information from these Cyrillic script documents naturally results in an underrepresentation of such information in scholarly data.

To pave the way for reducing this imbalance, we focus on the task of extracting structured information from bibliographic references found at the end of scholarly publications—commonly referred to as *citation field extraction* (CFE)—in Cyrillic script languages (see Figure 1). For this task, we introduce a data set of Cyrillic script references for training and evaluating CFE models. As Cyrillic publications usually contain both Cyrillic and English references, the data set contains a small portion (7%) of English references as well. The data set can be used in various scenarios, such as cross-lingual citation recommendation (Jiang et al. 2018) and analyzing the scientific landscape and scientific discourse independent of the used languages (Martín-Martín et al. 2021). To showcase the utility of our data set, we train several sequence labeling models on our data and evaluate them against a GROBID model retrained on the same data. Throughout the paper we refer to the reference string parsing module of GROBID as just "GROBID". To the best of our knowledge, we are the first to train a CFE model, more specifically BERT, specialized in Cyrillic script references.

Our contributions can be summarized as follows.

1. We introduce a large data set of labeled Cyrillic reference strings,[1] consisting of over 100,000 synthetically generated references and over 700 references that were manu-

[1] In the course of this work, we use the terms "reference string" and "citation string" interchangeably.

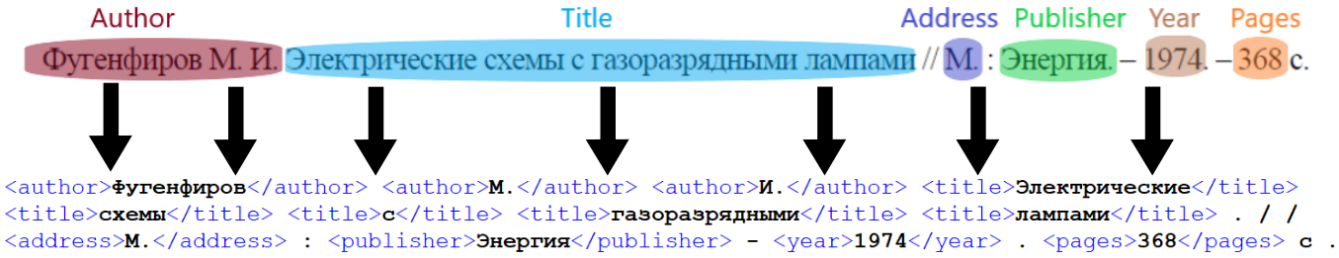

Figure 1: A real-world example of a Cyrillic script reference with marked bibliographic labels (top) and the corresponding labeled reference string (bottom).

ally labeled and gathered from multidisciplinary Cyrillic script publications.

2. We train the very first BERT-based citation field extraction (CFE) model specialized in Cyrillic script references and show the importance of retraining GROBID for Cyrillic script language data. We achieve an acceptably high F1 score of 0.933 with our best BERT model.

The data is available at https://doi.org/10.5281/zenodo.5801914, the code at https://github.com/igor261/Sequence-Labeling-for-Citation-Field-Extraction-from-Cyrillic-Script-References.

2 Related Work

CFE approaches that currently achieve the best performance are supervised machine learning approaches. Among them, the reference-parsing model of GROBID is typically reported to perform the best. We therefore use GROBID as the baseline in our evaluation.

In recent years, transformer-based models (Vaswani et al. 2017) such as BERT (Devlin et al. 2019) have achieved state-of-the-art evaluation results on a wide range of NLP tasks. To the best of our knowledge, there is so far only one paper presenting a BERT-based approach to CFE (Thai et al. 2020). The authors achieve state-of-the-art results on the UMass CFE data set (Anzaroot and McCallum 2013) by using ROBERTA, a BERT model with a modified training procedure and hyperparameters.

The original BERT model comes in three varieties, one trained on English text only, one on Chinese, and a multilingual model. Furthermore, many offshoots of BERT for different languages can be found in the literature. For Cyrillic languages, for example, RUBERT is a BERT variant trained on Russian text (Kuratov and Arkhipov 2019), and SLAVIC BERT is a named entity recognition model that was trained on four Slavic languages (Russian, Bulgarian, Czech, and Polish) (Arkhipov et al. 2019). Both of the aforementioned publications present a performance gain compared to the pretrained multilingual BERT by retraining on task-relevant languages. Because references in Cyrillic publications typically also contain a mix of Cyrillic and English references, we use multilingual BERT in our evaluation.

Table 1: A selection of existing citation data sets.

Data set	# Instances	Discipline
GROBID	6,835	Multi-discipline (Grobid's data set is a collection of various citation data sets)
CORA	1,877	Computer Science
UMass CFE	1,829	Science, technology, engineering, and mathematics
GIANT	911 million	Multi-discipline

3 Data Set

3.1 Existing Data Sets

Several publicly available data sets for training and evaluating CFE models exist. In Table 1, we show an overview of these citation data sets, including the number of reference strings contained and disciplines covered. In the following, we describe each of the data sets in more detail.

The authors of GROBID (Lopez 2009) provide the 6,835 samples their tool's reference parser is trained on. These are gathered from various sources (e.g., CORA, HAL archive, and arXiv). New data is continuously added to the GROBID data set[2].

One of the most widely used data sets for the CFE task is CORA,[3] which comprises 1,877 "coarse-grained" labeled instances from the computer science domain. As pointed out by Prasad, Kaur, and Kan (2018), a shortcoming of the CFE research field is that the models are evaluated mainly on the CORA data set, which lacks diversity in terms of multidisciplinarity and multilinguality.

The UMass CFE data set by Anzaroot and McCallum (2013) provides both fine- and coarse-grained labels from across the STEM fields. Fine- and coarse-grained labels means, for example, that labels are given for a person's full name (coarse-grained), but also for their given and family name separately (fine-grained).

[2] See https://github.com/kermitt2/grobid/issues/535.

[3] See https://people.cs.umass.edu/~mccallum/data.html.

All of the above manually annotated data sets are rather small and part of them is limited in terms of the scientific disciplines covered. These issues are addressed by Grennan et al. (2019) with the data set GIANT, created by synthetically generating reference strings. The data set consists of roughly 1 billion references from multiple disciplines, which were created using 677,000 bibliographic entries from Crossref[4] rendered in over 1,500 citation styles.

We see none of the data sets described above as suitable for training a model for extracting citation data from Cyrillic publications' references, because they are based on English language citation strings only, except for GIANT. However, GIANT does not provide consistent language labels, making the issue of accurate filtering for Cyrillic script citation strings non-trivial.

To the best of our knowledge, no data set of citation strings in Cyrillic script currently exists. It is therefore necessary to create a data set of labeled citation strings to be able to train models capable of reliably extracting information from Cyrillic script reference strings.

3.2 Data Set Creation

In the following subsection, we identify two approaches for creating an appropriate data set to train and test deep neural networks that extract citation fields, such as author information and paper titles. Grennan et al. (2019), Grennan and Beel (2020), and Thai et al. (2020) found that synthetically generated citation strings are suitable to train machine learning algorithms for CFE, resulting in high-performance models. We use a similar approach to create a synthetic data set of citation strings for model training in the next section. To evaluate the resulting models on citation strings from real documents, we manually annotate citation strings from several Cyrillic script scientific papers. This is described in the subsection "Manually Annotated References."

Synthetic References Figure 2 shows a schematic overview of our data set creation, which is described in the following.

To create a data set of synthetic citation strings, a suitable source of metadata of Cyrillic script documents is necessary. Crossref, which is used by GIANT, provides metadata for over 120 million records[5] of various content types (e.g., journal-article, book, and chapter) via their REST API. Unfortunately, most of the data either does not provide a language field or the language tag is English. We also considered CORE (Knoth and Zdrahal 2012) as a source of metadata. Although CORE provides at least 23,000 papers with Cyrillic script language labels and corresponding PDF files (Krause et al. 2021), it comes with insufficient metadata. Furthermore, for the relevant BibTeX fields, CORE only provides title, authors, year, and some publisher entries.

We identified Web of Science (WoS)[6] as the most appropriate source of metadata for creating synthetic references and based on the option to gather language-specific metadata. Additionally, WoS provides a filter for the document

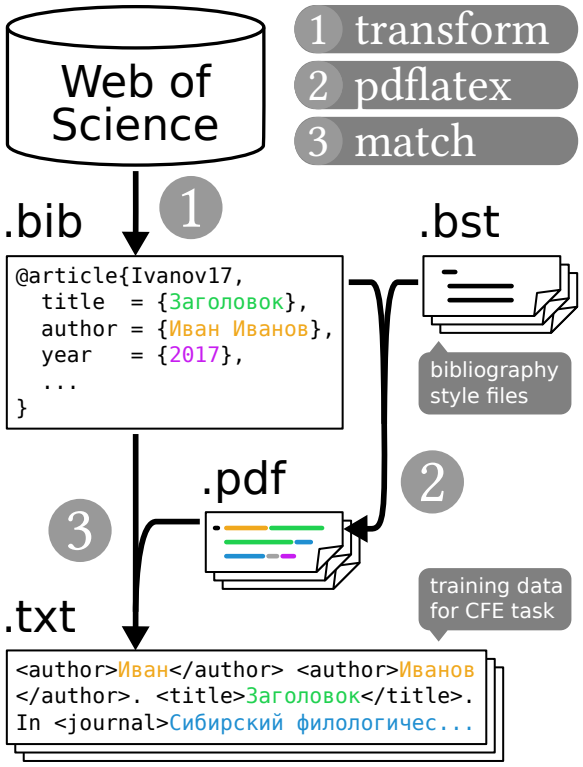

Figure 2: Schematic overview of the synthetic data set creation.

type, even though it lacks, for example, book types. The final data set should contain multiple document types to cover various citation fields.

Web of Science provides access to the Russian Science Citation Index (RSCI), a bibliographic database of scientific publications in Russian with roughly 750,000 instances. We chose to gather around 27,000 most recent (i.e., from 2020) article type and around 7,000 most recent (i.e., from 2010-2020) conference proceeding type[7] metadata records from the RSCI. The selection is motivated by the finding of Grennan and Beel (2020) that a model trained with more than 10,000 citations would decrease in performance compared with a smaller training data set. To verify the latter statement in our evaluation, we decide to create a data set consisting of 100,000 citation strings in total. Last but not least, following the GIANT data set, we wanted our data set to consist of around 80% articles and 20% conference proceedings.

Based on the language tags in the metadata provided by WoS, a breakdown of the languages of the data we collected is shown in Table 2. Unfortunately, the RSCI database by WoS does not provide Ukrainian language metadata, but since Russian and Ukrainian are very similar, we expect the model to process Ukrainian language references comparably reliable to Russian language references. In our evaluation, we show that our model achieves similar F1 scores for

[4]See https://www.crossref.org.

[5]See https://api.crossref.org/works.

[6]See https://www.webofknowledge.com/.

[7]The conference proceeding type corresponds to meeting type in WoS.

Table 2: Distribution of the reference languages from WoS.

Language	Number of items
Russian	31,977
English	2,241
other	9

Russian and Ukrainian language references.

After converting the WoS data to the BibTeX format and filtering out corrupted entries, we enrich the data with additional features, such as "Pagetotal"[8] and "address" (publisher city), to get extensive BibTeX entries that are comparable to real references. This process results in a total of 34,228 metadata records in the BibTeX format. To generate bibliographic references, we additionally need to identify a set of suitable citation styles.

Based on a CORE subset of Cyrillic script scientific papers (see next subsection for details), we identify the GOST and *APA* citation styles to be best suited for generating realistic reference strings. The GOST standards[9] were developed by the government of the Soviet Union and are comparable to standards by the American ANSI or German DIN. They are still widely used in Russia and in many former soviet republics. To introduce a certain level of variety we use the *GOST2003*, *GOST2006*,[10] and *GOST2008* styles for all references. Since the *APA* style cannot handle Cyrillic characters, it is used for non-Cyrillic references only.

For each reference, we create a separate PDF rendition. Using various bibliography styles for the same reference can result in reference strings that are completely different in look and structure. For instance, author names can be abbreviated or duplicated at different positions.[11] Metadata labels and their counterparts in the PDF references are then matched by an exact string match or, alternatively, the Levenshtein distance. Exact string matches are not always possible because some characters are manipulated by TeX while generating a PDF file or field values themselves change during the generation process in various ways, like abbreviations or misinterpreted characters. To store the reference text and reference token labels in one file per reference, we create labeled reference strings as shown in Figure 1.

In rare cases during the parsing process of the PDFs to text strings using *PDFMiner*, tokens were garbled and files could not be read. Consequently, the corresponding items are removed from the data set, resulting in slightly varying numbers of references for different citation styles. In the end, our approach yields about 100,000 synthetically generated labeled reference strings. A detailed breakdown of the quantity of data for each citation style is shown in Table 3.

In Table 4, we additionally show the breakdown of labels covered by our synthetic references.

Table 3: Number of synthetic labeled reference strings per citation style & reference type.

Citation Style	# Articles	# Conf. Proc.	Total
APA	1,293	833	2,126
GOST2003	26,289	7,061	33,350
GOST2006	26,328	7,078	33,406
GOST2008	26,467	7,113	33,580
Total	80,377	22,085	102,462

Table 4: Number of synthetic labeled reference strings having respective labels per reference type.

Label	# Articles	# Conf. Proc.	Total
title	80,376	22,085	102,461
author	80,375	22,079	102,454
year	80,305	21,870	102,175
pages	80,419	17,944	97,113
journal	80,376	–	80,376
number	80,214	–	80,214
volume	46,494	11,423	57,917
booktitle	–	22,085	22,085
publisher	–	22,083	22,083
address	–	20,034	20,034
pagetotal	1,208	4,141	5,349

Manually Annotated References Despite the fact that many large scholarly data sets are publicly available, most lack broad language coverage or do not contain full text documents. Investigating several data sources, we find that, for example, the PubMed Central Open Access Subset[12] provides mostly English language publications,[13] just like S2ORC (Lo et al. 2020). Further, the Microsoft Academic Graph (Sinha et al. 2015; Wang et al. 2019) covers millions of publications, but does not contain full texts and therefore also no reference strings.

We use the data set introduced by Krause et al. (2021) as a source of Cyrillic script papers. After a filtering step to remove papers with lacking or unstructured citations we randomly chose 100 papers to manually annotate.

Analyzing the origin of the selected papers, we note that 80 originate from the "A.N.Beketov KNUME Digital Repository"[14] and five from the "Zhytomyr State University Library."[15] Origins could not be determined for 15 papers. Figure 3 shows the distribution of papers by publication year. A breakdown of the disciplines covered by the data

[8] "Pagetotal" is a field specific to the citation style "GOST", which will be discussed later.

[9] See https://dic.academic.ru/dic.nsf/ruwiki/79269.

[10] Because we were not able to find a copy of the *GOST2006* BST file, we replicated it ourselves based on the *GOST2003* BST file and the description at https://science.kname.edu.ua/images/dok/journal/texnika/2021/2021.pdf.

[11] An example for a duplicated author name is shown in the following GOST2006 style reference: "Alefirov, A.N. Antitumoral effects of Aconitum soongaricum tincture on Ehrlich carcinoma in mice [Text] / Alefirov, A.N. and Bespalov, V.G. // Obzory po klinicheskoi farmakologii i lekarstvennoi terapii.–St. Petersburg : Limited Liability Company Eco-Vector.–2012.".

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

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

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

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

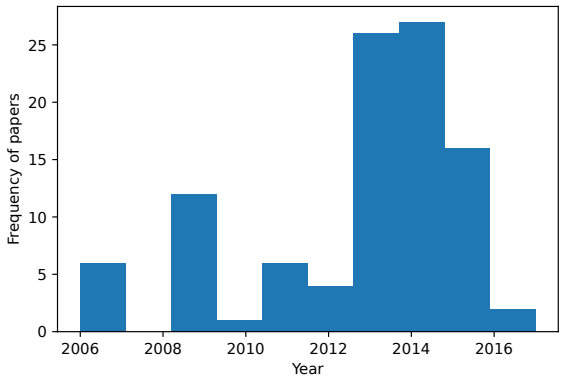

Figure 3: Distribution of publication years of the selected 100 papers.

set revealed that the most strongly represented disciplines are "engineering" with 36 papers and "economics" with 16 papers. The remaining 48 papers are spread across various fields, such as education, zoology, urban planning/infrastructure.

Using fastText (Joulin et al. 2016b,a) language detection, we find that our sample consists of 65 Ukrainian language and 35 Russian language papers.

Using the annotation tool INCEpTION (Klie et al. 2018), we label the references in our 100 PDFs. Regarding manual annotation, we note that the real references did not always fit our set of metadata labels. For example, references to patents, legal texts, or web resources might not contain certain elements typical for references to scientific papers. Furthermore, references containing fields outside the scope of our labels, like *editor* or *institution*, exist. In the case of *booktitle* fields of conference proceedings, we used the *journal* label. Lastly, due to the difference in use of "№" across citation styles (indicating either an issue or volume number), in ambiguous cases the number after "№" is labeled *volume* following the *GOST2006* citation style.

Table 5 shows the summary statistics of the resulting data set. In Table 6, we show the labels used and their number of occurrences counted in segments (a segment is the full text range for a label).

Although 65% of the 100 documents are Ukrainian language papers, the references are written in various languages. Nearly 99% are written in Russian, Ukrainian and English (see Table 7). Other languages contained are Polish, German, Serbian, and French.

While the number of manually annotated references is not large enough for training purposes, we argue that the size and language distribution enable us to perform a realistic evaluation of our models.

4 Approach

There are various approaches to the CFE task. Most of them use regular expressions, template matching, knowledge bases, or supervised machine learning, whereby machine learning-based approaches achieve the best results

Table 5: Summary of the manually annotated data set.

Parameter	Counts
Number of annotated papers	100
Number of reference strings	771
Average reference length (in tokens)	28.00
Number of reference related labels	11
Number of labeled reference segments	5,080

Table 6: Segment counts for the labels assigned.

Label	#segments
author	1,560
title	773
year	680
pages	612
address	410
publisher	364
journal	328
volume	256
number	91

(Tkaczyk et al. 2018). Furthermore, tools differ in terms of extracted reference fields and their granularity.

GROBID is commonly considered as the most effective tool (Tkaczyk et al. 2018) and was created by Lopez. Tkaczyk et al. reported an F1 score of 0.92 for the retrained GROBID CRF model on their data set. Beyond parsing reference strings, GROBID is also able to extract metadata and logical structure from scientific documents in PDF format.

Following existing literature, we decide to use the GROBID CRF model as a baseline. Therefore we retrain the GROBID CRF model on our synthetic data set following GROBID's documentation.[16] The GROBID CRF model is trained from scratch.[17]

State-of-the-art sequence labeling approaches are often based on BERT. Accordingly, we fine-tune the cased multilingual BERT model, which is pretrained on 104 languages, on our synthetic reference data set. We fine-tune/retrain both BERT and GROBID on several subsets of our synthetic data set with differing sizes (between 500 and 100K) so that we can assess the necessity of a large training set.

[16] See https://grobid.readthedocs.io/en/latest/Training-the-models-of-Grobid/.

[17] See https://github.com/kermitt2/grobid/issues/748.

Table 7: Distribution of the reference languages in the manually annotated data set.

Language	Number of references
Russian	390
Ukrainian	288
English	82

Table 8: Evaluation on manually annotated data set for BERT models with differing sizes of training data average over 5 models trained on different random samples.

Train Set Size	Recall	Precision	F1 Score	Standard Deviation
500	0.909	0.916	0.910	0.007
1,000	0.922	0.926	0.923	0.009
2,000	**0.928**	**0.932**	**0.928**	0.007
3,000	**0.928**	0.931	**0.928**	0.003
5,000	0.926	0.929	0.927	0.004
10,000	0.920	0.925	0.921	0.005
20,000	0.907	0.913	0.907	0.008
50,000	0.863	0.880	0.864	0.017
100,000	0.847	0.868	0.848	0.012

5 Evaluation

Fine-tuning the BERT model is, compared to pretraining, relatively inexpensive (Devlin et al. 2019). We observed this as well by comparing the time for fine-tuning with the time needed to train GROBID. For example, fine-tuning BERT with 100,000 training instances takes 125 minutes (on a GeForce RTX 3090 GPU) and training GROBID CRF (on a 16 core Intel Xeon Gold 6226R 2.90GHz CPU) takes 1,233 minutes.

To evaluate our fine-tuned BERT model not only on the manually annotated but also on the synthetic references, we remove a hold-out set of 2,000 synthetic references from the training set, with a fixed distribution of citation styles, according to the distribution of the entire data set.

5.1 BERT Evaluation on the Manually Annotated Data Set

We fine-tune the cased multilingual BERT model on 9 training set sizes from our synthetically generated labeled reference data. To ensure robust results, for each of the 9 training set sizes, we sample 5 training sets, train one model per sample and average the resulting scores (i.e., in total we train $9 \times 5 = 45$ models).

Averaged scores for recall, precision, and F1 score for all 9 training set sizes are visualized in Figure 8.We found that models trained on relatively small training data sets (between 1,000 and 10,000 instances) perform best on our manually annotated test set. More precisely, on average, the models trained on 2,000 instances perform best regarding the F1 score. These models achieve an average F1 score of 0.928 (range from 0.917 to 0.936). Already with the smallest considered training set of 500 instances, we can fine-tune a powerful BERT model for the Cyrillic CFE task achieving an F1 score of 0.91 on average.

The highest achieved F1 score of 0.928 (averaged F1 scores of five models trained on different 2,000 instances random samples) on our test set is comparable with state-of-the-art models proposed for English CFE (Tkaczyk et al. 2018; Grennan and Beel 2020; Thai et al. 2020; Prasad, Kaur, and Kan 2018), especially considering the fact that there are reference types and languages in the test set the model was not trained on. Nevertheless, it is difficult to compare our results with other papers, since we work with Cyrillic script references and evaluate the models on our self-created test set.

Table 9: Detailed evaluation of labels predicted by BERT_{Final}.

Label	Prec.	Rec.	F1	Supp.
author	0.984	0.994	0.989	7,104
year	0.945	0.962	0.953	680
pages	0.922	0.984	0.952	1,112
address	0.927	0.961	0.944	715
other	0.945	0.926	0.936	10,730
title	0.938	0.931	0.934	7,257
publisher	0.913	0.781	0.842	1,165
journal	0.765	0.861	0.810	1,982
volume	0.836	0.454	0.588	269
number	0.345	0.860	0.492	93
Weighted Average Score	0.936	0.932	0.933	31,107

We further evaluate a BERT model trained on 2,000 random instances[18]—referred to as BERT_{Final} from here on—regarding individual labels. Since our model is more fine-grained than the test set, i.e. labels in the synthetic data set and manually annotated data set are not the same, we had to change the *pagetotal* label to *pages* and the *booktitle* label to *journal*.

As shown in Table 9, our model performs best on identifying author tokens with an F1 score of 0.989. Overall, we observe an F1 score of more than 0.934 for 6 labels (*author, year, pages, address, other*, and *title*).

We see room for improvement in *publisher*, *journal*, *volume*, and *number* predictions. The poor performance in *volume* and *number* predictions can be explained by the ambiguity of "№" in the test set (see Section "Manually Annotated References").

We see high recall with low precision values in *number* predictions and low recall with high precision values in *volume* predictions. The same observation can be made for *journal* and *publisher* predictions, but to a lesser degree.

More than 50% of the actual *volume* labels are labeled as *number*, and around 17% of actual *publisher* labels are labeled as *journal*.

Next, we look into the evaluation on the synthetic hold-out set. We evaluate the BERT_{Final} model depending on the languages of references (see Figure 4).

As mentioned before, our synthetic data set lacks Ukrainian language references. Nevertheless, the F1 score of 0.946 for Russian language references is only 2.5% higher than the F1 score of 0.921 for Ukrainian language references. This is potentially due to the high similarity between the Russian and Ukrainian languages.

Additionally, for English language references, the predictions of *volume* and *number* labels are much better than for

[18]Models trained on 2,000 instances perform best on average.

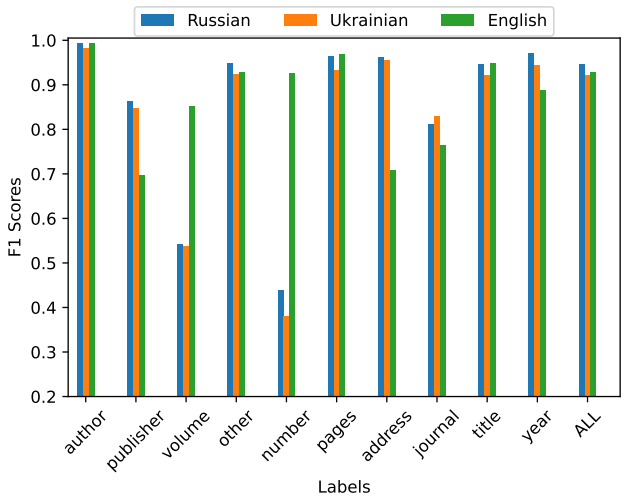

Figure 4: Evaluation on manually annotated data set for BERT$_{Final}$ model per label and language

Cyrillic script references. This is due to the fact that most English language references are formatted in the *APA* style, where there is no ambiguity in the respective labels.

Furthermore, BERT$_{Final}$ predicts *publisher* and *address* labels worse for English language references than for Russian and Ukrainian language references.

5.2 BERT Evaluation on the Synthetic Hold-Out Set

Our fine-tuned BERT underperforms in some labels on the manually annotated test set. To evaluate our model on data with less ambiguity and the same reference document types it was trained on, we assess the performance on the synthetic hold-out set.

Scores for recall, precision, and F1 score for all 9 training set sizes evaluated on the hold-out set are visualized in Figure 5. All BERT models achieve F1 scores of over 0.99, even the model fine-tuned with 500 instances. We also see a steady increase in the performance, when increasing the training data set size. Best performance regarding the F1 score (0.998) is achieved by the model trained on 100,000 instances, while this model performs worst on the manually annotated data set. There are also small differences in the scores concerning individual labels.

5.3 GROBID Evaluation

We compare our fine-tuned BERT with the state-of-the-art GROBID model. First, we evaluate the off-the-shelf GROBID on our manually annotated test set. The model achieves unsatisfying results with an F1 score of 0.09. Only numeric tokens such as *number* or *year* achieve an F1 score of over 0.1. Most of the non-numeric labels have a F1 score of 0 or close to 0.[19]

[19]Data used for training of the off-the-shelf GROBID has different labels than we have in our synthetic data set. Consequently some labels are condemned to have scores equal zero, e.g. *web*.

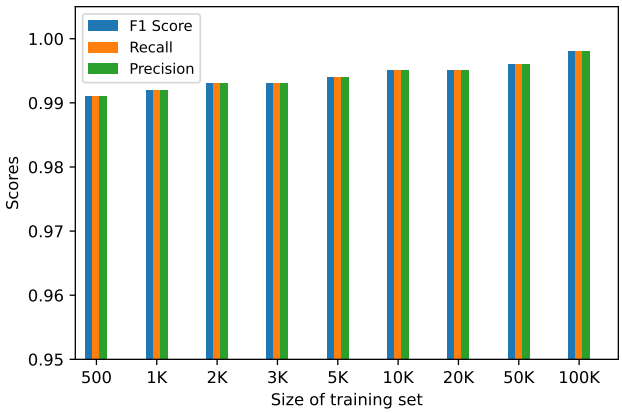

Figure 5: Evaluation on synthetic hold-out data set for BERT models with differing size of training data

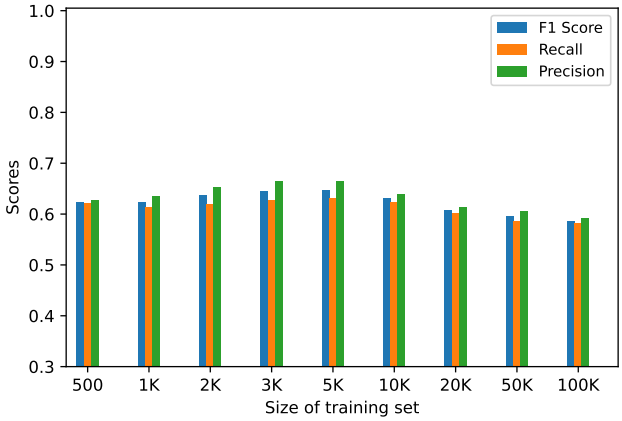

Figure 6: Evaluation on real data set for GROBID CRF models with differing sizes of training data sets.

GROBID was initially trained on English language references. Consequently, it is not surprising that it performs poorly regarding Cyrillic reference data. Therefore, we retrain the GROBID CRF model on our synthetic Cyrillic reference data with differing training data set sizes, as we did for the BERT model. Evaluations of resulting models on our manually annotated test set are shown in Figure 6.

We observe poorer performance of the GROBID models compared to our fine-tuned BERT. Similar to evaluations of the fine-tuned BERT models and Grennan and Beel (2020), we see that the best performing models where trained on relatively small data sets consisting of less than 10,000 references. The best performing GROBID model was trained with 5,000 instances, achieving a F1 score of 0.647. We refer to this best performing GROBID model as GROBID$_{Final}$. Compared to the off-the-shelf GROBID results, we managed to increase the F1 score by a factor of seven by retraining GROBID.

Note that GROBID does not provide evaluation scores for *other* labels.

Table 10: Summary of metrics of the models evaluated on the manually annotated test set.

Model	Precision	Recall	F1 Score
Vanilla GROBID	0.347	0.052	0.090
GROBID_{Final}	0.665	0.631	0.647
BERT_{Final}	**0.936**	**0.932**	**0.933**

Compared to the off-the-shelf GROBID, we see higher F1 scores in almost every label, except for *year* and *number*. The best label performance is measured for paper title, with an F1 score of 0.817. A comparison of evaluation metrics of GROBID and BERT is shown in Table 10. Our BERT_{Final} model outperforms the GROBID_{Final} model in every label and, consequently, in overall F1 score as well.

6 Conclusion

In this paper, we provide a large data set covering over 100,000 labeled reference strings in various citation styles and languages, of which 771 are manually annotated references from 100 Cyrillic script scientific papers. Furthermore, we fine-tune multilingual BERT models on various training set sizes and achieve the best F1 score of 0.933 with 2,000 training instances. We show the eligibility of synthetically created data for training CFE models. To compare our results with existing models, we retrained a GROBID model serving as a benchmark. Our BERT model significantly outperformed both off-the-shelf and retrained GROBID. In future work, our BERT model could be compared to other well-performing CFE models, such as CERMINE and NEURAL PARSCIT.

Our data sets can be reused by other researchers to train Cyrillic script CFE models. In particular our manually annotated data set can serve as a benchmark for further research in this field, since it provides references from various domains and covers several languages.

Regarding our BERT model, we see two key aspects for future work. First, literature describes benefits of adding a CRF layer at the top of a model's underlying architecture (Prasad, Kaur, and Kan 2018; Arkhipov et al. 2019), which could also be considered for our approach. Second, our model's performance could be increased by retraining BERT from scratch on task-specific languages, e.g. in our case Cyrillic Script languages and English, as shown by Kuratov and Arkhipov (2019) and Arkhipov et al. (2019).

References

Anzaroot, S.; and McCallum, A. 2013. A New Dataset for Fine-Grained Citation Field Extraction. *ICML Workshop on Peer Reviewed and Publishing Models*.

Arkhipov, M.; Trofimova, M.; Kuratov, Y.; and Sorokin, A. 2019. Tuning Multilingual Transformers for Language-Specific Named Entity Recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, 89–93. Florence, Italy: Association for Computational Linguistics.

Beel, J.; Gipp, B.; Langer, S.; and Breitinger, C. 2016. Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17(4): 305–338.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv pre-print server*.

Färber, M.; and Jatowt, A. 2020. Citation recommendation: approaches and datasets. *Int. J. Digit. Libr.*, 21(4): 375–405.

Grennan, M.; and Beel, J. 2020. Synthetic vs. Real Reference Strings for Citation Parsing, and the Importance of Retraining and Out-Of-Sample Data for Meaningful Evaluations: Experiments with GROBID, GIANT and Cora. *ArXiv*, abs/2004.10410.

Grennan, M.; Schibel, M.; Collins, A.; and Beel, J. 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*, 101–112.

Jiang, Z.; Yin, Y.; Gao, L.; Lu, Y.; and Liu, X. 2018. Cross-language Citation Recommendation via Hierarchical Representation Learning on Heterogeneous Graph. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR'18, 635–644.

Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; and Mikolov, T. 2016a. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2016b. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759*.

Khabsa, M.; and Giles, C. L. 2014. The Number of Scholarly Documents on the Public Web. *PLoS ONE*, 9(5): e93949.

Klie, J.-C.; Bugert, M.; Boullosa, B.; Eckart de Castilho, R.; and Gurevych, I. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5–9. Association for Computational Linguistics.

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

Krause, J.; Shapiro, I.; Saier, T.; and Färber, M. 2021. Bootstrapping Multilingual Metadata Extraction: A Showcase in Cyrillic. In *Proceedings of the Second Workshop on Scholarly Document Processing*, 66–72.

Kuratov, Y.; and Arkhipov, M. 2019. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. *arXiv pre-print server*.

Lo, K.; Wang, L. L.; Neumann, M.; Kinney, R.; and Weld, D. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4969–4983. Association for Computational Linguistics.

Lopez, P. 2009. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In Agosti, M.; Borbinha, J.; Kapidakis, S.; Papatheodorou, C.; and Tsakonas, G., eds., *Research*

and Advanced Technology for Digital Libraries, 473–474. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-04346-8.

Martín-Martín, A.; Thelwall, M.; Orduña-Malea, E.; and López-Cózar, E. D. 2021. Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. *Scientometrics*, 126(1): 871–906.

Moskaleva, O.; Pislyakov, V.; Sterligov, I.; Akoev, M.; and Shabanova, S. 2018. Russian Index of Science Citation: Overview and review. *Scientometrics*, 116(1): 449–462.

Ortega, J. L. 2014. *Academic search engines: A quantitative outlook*. Elsevier. ISBN 1780634722.

Prasad, A.; Kaur, M.; and Kan, M.-Y. 2018. Neural ParsCit: a deep learning-based reference string parser. *International Journal on Digital Libraries*, 19(4): 323–337.

Shaw, W. 1981. Information theory and scientific communication. *Scientometrics*, 3(3): 235–249.

Sinha, A.; Shen, Z.; Song, Y.; Ma, H.; Eide, D.; Hsu, B.-J. P.; and Wang, K. 2015. An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, 243–246. ACM. ISBN 978-1-4503-3473-0.

Thai, D.; Xu, Z.; Monath, N.; Veytsman, B.; and McCallum, A. 2020. Using BibTeX to Automatically Generate Labeled Data for Citation Field Extraction. In *Automated Knowledge Base Construction*.

Tkaczyk, D.; Collins, A.; Sheridan, P.; and Beel, J. 2018. Machine Learning vs. Rules and Out-of-the-Box vs. Retrained: An Evaluation of Open-Source Bibliographic Reference and Citation Parsers. *arXiv pre-print server*.

Tkaczyk, D.; Szostek, P.; Fedoryszak, M.; Dendek, P. J.; and Bolikowski, L. 2015. CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(4): 317–335.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Aidan; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. *arXiv pre-print server*.

Wang, K.; Shen, Z.; Huang, C.; Wu, C.-H.; Eide, D.; Dong, Y.; Qian, J.; Kanakia, A.; Chen, A.; and Rogahn, R. 2019. A Review of Microsoft Academic Services for Science of Science Studies. *Frontiers in Big Data*, 2: 45.