Vision and Natural Language for Metadata Extraction from Scientific PDF Documents: A Multimodal Approach

Conference Paper · April 2022

DOI: 10.1145/3529372.3533295

CITATIONS
2

READS
434

2 authors:

Zeyd Boukhers
Fraunhofer Institute for Applied Information Technology FIT
51 PUBLICATIONS **157** CITATIONS

SEE PROFILE

Azeddine Bouabdallah
Universität Koblenz-Landau
4 PUBLICATIONS **2** CITATIONS

SEE PROFILE

Vision and Natural Language for Metadata Extraction from Scientific PDF Documents: A Multimodal Approach

Zeyd Boukhers and Azeddine Bouabdallah
University of Koblenz-Landau, Germany
{boukhers,bazeddine}@uni-koblenz.de

ABSTRACT

The challenge of automatically extracting metadata from scientific PDF documents varies depending on the diversity of layouts within the PDF collection. In some disciplines such as German social sciences, the authors are not required to generate their papers according to a specific template and they often create their own templates which yield a high appearance diversity across publications. Overcoming this diversity using only Natural Language Processing (NLP) approaches is not always effective which is reflected in the metadata unavailability of a large portion of German social science publications. Therefore, we propose in this paper a multimodal neural network model that employs NLP together with Computer Vision (CV) for metadata extraction from scientific PDF documents. The aim is to benefit from both modalities to increase the overall accuracy of metadata extraction. The extensive experiments of the proposed model on around 8800 documents proved its effectiveness over unimodal models, with an overall F1 score of 92.3%.

CCS CONCEPTS

• **Information systems** → *Data encoding and canonicalization*; • **Computing methodologies** → **Supervised learning by classification**.

KEYWORDS

metadata extraction, multimodal ML, NLP, CV

1 INTRODUCTION

With the continuous expansion usage of digital libraries, a huge amount of scientific papers are published every year in a digital format. Specifically, nearly two million scientific papers are published each year [2]. These papers require automatic processing to ease their use for scholars such as querying papers, citation count, paper recommendations, etc. Therefore, the availability of metadata (i.e. title, authors, year of publication, etc.) is important. However, in some disciplines such as German social science, an important number of the published papers are not covered in accessible bibliographic databases [9, 26]. This means that the metadata can only be obtained by extracting it from PDF documents.

Intuitively, the metadata is extracted using NLP approaches [21], which demonstrated their efficiency on English documents due to the relatively standard layout in English corpora [6]. German scientific papers often come in a large variety of layouts because they are mainly published by small and mid-size publishers who do not impel to use standard templates. Since the German social science community is relatively small, there is not much work addressing this problem [10, 12]. To overcome this, we proposed in an earlier study [6] to tackle this problem using CV techniques by viewing the PDF as an RGB image. Although this approach demonstrated promising performance on a challenging dataset, it still fails to accurately extract some patterns (e.g. DOI), which are supposed to be easily extracted by NLP-based approaches.

Introducing and considering multiple types of input data by combining NLP and CV has proven its effectiveness in previous works in different fields [28]. Therefore, this paper tackles the problem of automatically extracting metadata from scientific documents in the German language using a multimodal approach that views a PDF document both as an RGB image and as a textual document. With this, we assume that jointly learning both modalities can lead to a better understanding of the documents.

To this end, we trained a multimodal neural network model with two sub-models; the first one is a BiLSTM model fed with the layout and context features of the content and the second one takes as input the image representation of the PDF document. Using late fusion, the output vectors generated by the two sub-models are concatenated and used as input to another BiLSTM model which classifies each token.

Following this section, Section 2 discusses the related works. Section 3 presents the proposed approach and Section 4 presents the conducted experiments and the obtained results that validate the effectiveness of the proposed approach. Finally, Section 5 concludes this paper and gives insight into future directions.

2 RELATED WORK

In this section, we categorize the related works on extracting metadata from PDF documents into three main categories.

2.1 Natural Language Processing

[14] considers two types of metadata extraction methods, namely machine learning-based [15, 23, 24] and rule-based approaches [16, 19]. Machine learning approaches such as CiteSeerX [20] train models on labelled datasets and then apply them to extract metadata from new data. Examples of these models include Hidden Markov Models (HMM) [24], Conditional Random Fields (CRFs) [23] and Support Vector Machines (SVM) [15]. Although machine learning techniques are considered to be robust and effective, labelling the training dataset can be a time-consuming process, especially in tasks with a huge diversity among samples. Contrary to machine learning, rule-based approaches use a set of defined rules guiding how to extract metadata based on human observation [14].

In a few recent works, the problem of metadata extraction has been tackled using Deep Neural Networks (DNN), taking advantage of their enormous development. The obtained results proved that DNN significantly outperforms traditional methods [14]. [18] proposed a Bidirectional LSTM-CRF model to encode a sequence of word representations, and a CRF layer to extract the predicted label

sequence. [11] used a Bidirectional LSTM along with a Convolutional Neural Network (CNN) to compute character-level word representation. [3] introduced a metadata extraction via DNN-based Segment Sequence Labeling outperforming existing works like ParsCit [13], an open-source CRF-based model, and BibPro [8], a neural network-based model, on the public datasets UMass [4] and Cora [24].

2.2 Computer Vision

CV-based approaches have yet to be widespread across the field of metadata extracting, however, in many recent works, they achieved promising results when tackling NLP-related problems. For instance, DeepPDF [25] is a method for segmenting paragraphs of a PDF document by viewing the document as an image and using UNet-Zoo, an architecture for biomedical image segmentation, to identify the paragraph while ignoring the header, caption, figure and references. The goal was to demonstrate the proof of concept that CV-based approaches are capable of analysing textual documents.

Inspired by [25], MexPub [6] proposed a method to extract metadata from German PDF documents at a pixel-by-pixel level. To this end, MexPub used MASK-RCNN architecture [17], which is dedicated to object detection and classification using ResNeXt [29] as backbone and Feature Pyramid Networks (FPN) to extract features from raw images. Although the results of MexPub are promising, the method still faces challenges such as the low generalization to scientific literature that has a significantly different structure from the trained dataset. Also, MexPub has still difficulties in precisely detecting some patterns when they are small and displayed in a different or uncommon position. Therefore, the results of MexPub can be improved by incorporating text processing in a joint architecture.

2.3 Multimodality

Multimodal deep learning has been used in several applications including audiovisual and image classification demonstrating a promising performance. In metadata extraction, in particular, multimodality has shown better results in comparison to unimodal counterparts as indicated in [5] and [22]. [5] considered both audio and video modalities to extract metadata from video lectures using a Naive Bayes classifier and a rule-based refiner. The approach relies on the correlations between the audio transcripts and the content in the slides embedded in the video streams. This combination has shown an improvement of 114.2% in terms of F-score, precision and recall compared to audio-based approaches.

[22] introduced a deep learning approach to extract metadata from scientific documents. This multimodal approach can process both image data and text data as sources of information and does not need to design any classification feature. The approach handles the text source using Recurrent Neural Networks (RNNs) and the image source with Convolutional Neural Networks (CNNs). Finally, the joint representation is fed into a BiLSTM network and the final classification is performed by a CRF classifier. The results of this combination proved its effectiveness over unimodal approaches. Despite this improvement, we assume that the model is prone to overfitting as the CV-based model is trained only with a few scientific documents which are not supposed to be sufficient to train the high number of parameters of CNN layers.

3 METHODOLOGY

To extract metadata considering both text and image modalities, we propose a Neural Network architecture that can be decomposed into three sub-models: NLP model, CV model and a classifier as illustrated in Figure 1. Given a PDF document, the text is extracted only from the first page using CERMINE [27] as it is one of the most reliable tools to extract texts from different layouts on the line level. The extracted lines are associated with information about the geometric structure such as text position and font style.

From each token, a set of 16 handcrafted features are extracted and concatenated with word embedding that encodes the context and the meaning of the words into feature vectors. The concatenated vector serves as input for the NLP sub-model. The CV model is fed with the image of the first page of the document.

NLP sub-model. Following [30], we employed a BiDirectional Long-Short-Term Memory (BiLSTM) to model the extracted text due to its validated efficiency. Consequently, our NLP sub-model is a two layers-LSTM model with 256 hidden dimensions each, where the first one is a forward LSTM and the second is a backward LSTM. The model takes as input a word embedding vector of length 1041 at each time step. As mentioned above, the embedding vector is composed of two parts, the first part of length 16 contains layout features (font size of the word, font style, spacing between the word and the line above/under, flags describing whether the text is in italic, bold or follows a specific common format such as date or email, etc.). The second part comprises the ELMO [7] embedding results obtained from a model trained on German documents. The two LSTM layers are followed by a fully connected layer of length 512 and an output layer of 10 neurones with a softmax activation function to obtain probability scores for the word belonging to each of the 10 classes (abstract, author, email, address, date, journal, affiliation, DOI, title, unclassified).

CV sub-model. Due to the effectiveness of our earlier work MexPub [6], we employ it as the CV sub-model fed with a 296x794 pixels image representing the first page of the given PDF document. The MexPub model is already trained on 30K generated German scientific papers from the SSOAR repository. In addition, MexPub uses transfer learning by re-training the model from PubLayNet dataset [31], which has been already trained on a large scientific literature dataset for classifying classes such as title, text, list, table, and figure and is later fine-tuned to extract metadata classes. Afterwards, we extract the text inside the bounding boxes given by the CV sub-model and then aggregate the probabilities for all possible classes in that box before feeding them into the classifier. Note that the CV sub-model might also output unclassified bounding boxes (i.e. it does not belong to any of the defined classes).

Classifier. At the end of our architecture pipeline, a SoftMax classifier is used, which fuses the output of the NLP and CV sub-models. Specifically, we extract all the words from the document and go through them sequentially and concatenate their vector representations, generated by the sub-models. Here also, we employed a

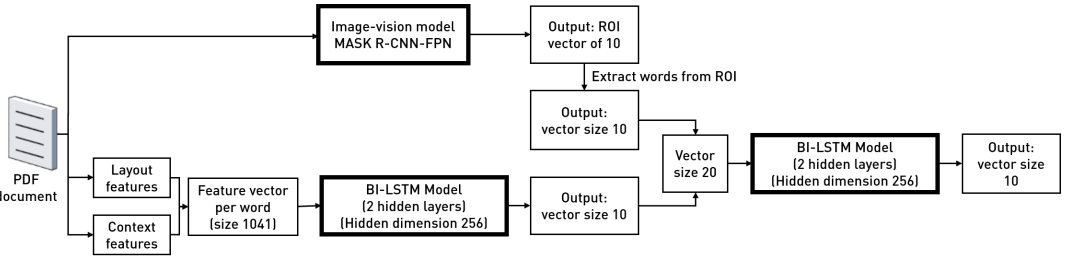

Figure 1: An overview of the proposed multi-modal architecture for metadata extraction

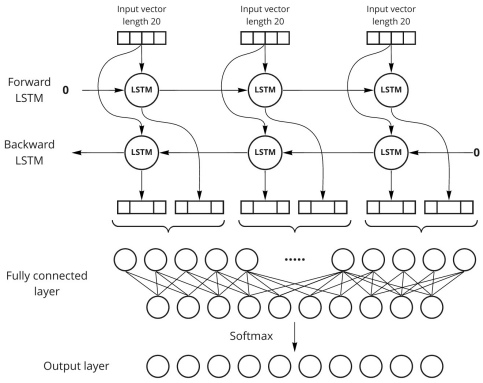

Figure 2: The proposed architecture of the classifier

BiLSTM is able to train in both directions and preserve information in both past and future states, which is particularly useful in recognizing the context and learning patterns within sentences or paragraphs (if the next and previous words are title, it is highly likely the current one is a title as well). Specifically, the model takes a vector of length 20, which is the concatenation of both outputs sub-models. The output of this model is also a probability distribution of length 10 corresponding to all classes. As Figure 2 illustrates, the classifier consists of two stacked LSTM layers (forward and backwards LSTMs) with 256 hidden dimensions. These two layers are followed by a fully connected layer containing 512 input nodes and 10 output nodes with a SoftMax activation function.

4 EXPERIMENTS AND RESULTS

Data. To validate the effectiveness of the proposed approach, it is important to collect a dataset of German publications that is labelled both on text and image levels. However, this is challenging as there is no available dataset and manual labelling is very time-consuming. Therefore, we manually annotated only 300 documents randomly selected from the SSOAR repository, which consists of German scientific publications from the field of Social Science. Based on the 13 different templates derived from the manually annotated documents, 8518 documents were generated using metadata records of German publications retrieved from SSOAR [1], MediaRep [2] and DBLP [3]. Note that some of these layouts have been identified and used previously [6] as well. We generated a document by inserting the collected metadata at their respective positions in the template.

[1]https://www.gesis.org/ssoar/home
[2]https://www.mediarep.org/
[3]https://dblp.org/xml/release/

To train and evaluate our model, we randomly split our training data into 70% training, 15% validation, and 15% testing. Due to the high complexity of the proposed architecture, each of the sub-models is trained independently using the extracted data. This allowed us to evaluate and improve each of the sub-models based on their results, making debugging easier and training faster. We trained each of the biLSTM models with 300 iterations and a batch size of 2000.

Results. The proposed approach achieved an overall F1-Score of 92.18% when training all the sub-models independently with 300 iterations. This validates the model's proficiency in utilizing contextual and visual features to accurately extract metadata from German scientific publications. Table 1 presents the results of both sub-models and the final multimodal one which combines the output of both of them. As demonstrated, the final model outperforms both NLP and CV sub-models. The table also shows that all models are effective at extracting large patterns such as "*Abstract*" compared to small ones (e.g. "*Date*"). Along with these promising results, the multimodal model achieved lower results compared to unimodal models for the class "*Affiliation*". We hypothesize that the NLP sub model was able to have higher results than the multimodal model because of its ability to recognize the patterns within the context of the text. But when adding the second modality, the computer vision model finds it hard to recognize due to the low occurrences and low variety of this class within the used dataset. Consequently, achieving lower results that affect the overall performance when combined with the NLP submodel.

Comparison against baseline approaches . To have a better understanding of the performance of our model, we compared the approach against other state-of-the-art approaches, namely GROBID [1] and MexPub [6]. For a fair comparison, we selected only English documents (300) from our testing set that our model has not seen during training. The reason is that Grobid is trained on English papers only and retraining it on German publications is a complex task. For all methods, we extracted the metadata from the given 300 documents and calculated the Cosine similarity between the extracted metadata and the ground truth of the document. If the cosine similarity is higher than a threshold of 0.85, the extracted metadata is considered correct. The reason for allowing a dissimilarity of 0.15 is the existence of cases where the extracted metadata misses a part of the text due to a deformation in the layout, but this does not necessarily mean the extraction is completely wrong. We compared the three methods in terms of F1-Score. Note that we did not include the evaluation for the classes such as "*Date*" and "*DOI*" because GROBID is not designed to extract them. Table 2

	NLP sub-model			CV sub-model			Final model		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Overall	0.835	0.819	0.827	0.908	0.898	0.904	**0.944**	**0.902**	**0.923**
Abstract	0.934	0.922	0.928	0.961	0.913	0.936	**0.989**	**0.962**	**0.975**
Author	0.860	0.960	0.907	0.932	0.940	0.936	**0.962**	**0.984**	**0.973**
Email	0.938	0.949	0.943	0.918	0.891	0.904	**0.985**	**0.953**	**0.969**
Address	0.631	0.857	0.727	0.897	0.906	0.901	**0.922**	**0.940**	**0.931**
Date	**0.980**	0.800	0.881	0.934	0.931	0.932	0.968	**0.972**	**0.970**
Journal	0.900	0.737	0.810	0.960	0.895	0.926	**0.971**	**0.937**	**0.954**
Affiliation	**0.820**	0.727	**0.771**	0.691	**0.742**	0.716	0.750	0.429	0.546
DOI	0.503	0.500	0.501	0.931	0.952	0.941	**0.986**	**0.967**	**0.976**
Title	0.956	0.919	0.937	0.950	0.920	0.935	**0.963**	**0.961**	**0.962**

Table 1: Performance comparison between the NLP and CV sub-models and the final multimodal model.

illustrates the results of the three models, demonstrating that ours outperforms both approaches despite the fact that the model does not strongly benefit from the NLP sub-model as it is well trained on German publications. Therefore, the results of our method and MexPub are comparable. The results of MexPub here are slightly different from the paper presented in [6] because the documents selected to evaluate all models are newly generated.

We hypothesize that the low results in the author and email classes compared to the ones from MexPub were due to the influence of the two sub-models. If only one of the models outputs a bad result, then the last model is going to be influenced badly by this. As a result, we presume that implementing a way to add a confidence level to each of the sub-models outputs would enhance the performance of the approach. By doing so, the last model will be able to weigh each sub-model output depending on its confidence level to prevent the influence of wrong results on the overall performance. Moreover, decreasing the model complexity could also help to solve the issue of error aggregation.

To validate the hypothesis that our multimodal model outperforms MexPub, we conducted another experiment using all test samples and compared the results of both models. Table 3 illustrates the results of our final model, our NLP sub model and MexPub in terms of Precision, demonstrating that the multimodal model outperforms both MexPub and the NLP sub-model.

	Our Model	MexPub	Grobid
Overall	**0.846**	0.823	0.618
Abstract	**0.923**	0.910	0.821
Author	0.807	**0.824**	0.770
Email	0.844	**0.901**	0.624
Address	**0.870**	0.821	0.324
Journal	**0.835**	0.828	0.741
Affiliation	**0.679**	0.535	0.240
Title	**0.964**	0.942	0.812

Table 2: Comparison of F1-Score achieved by our model, MexPub [6] and Grobid [1]

5 CONCLUSION

In this paper, we introduced a multimodal architecture to extract metadata from German scientific documents. The proposed model takes as input the textual content, the layout features and the image representation of the PDF document. The conducted experiments

	Our Model	MexPub	NLP Sub-model
Overall	**0.944**	0901	0.835
Abstract	**0.989**	0.957	0.956
Author	0.962	**0.975**	0.934
Email	**0.985**	0.917	0.860
Address	**0.922**	0.879	0.900
Date	**0.968**	0.892	0.631
Journal	**0.971**	0.859	0.938
Affiliation	0.975	0.809	**0.980**
DOI	**0.986**	0.945	0.500
Title	**0.963**	0.876	0.820

Table 3: Comparison of precision achieved by our multimodal model, unimodal model (NLP) and MexPub [6]

prove the outperformance of this multimodal model over unimodal ones.

For future work, we aim to improve the model by considering both early and late fusion to let the model fully benefit from multimodality. We will also use a CRF classifier as it demonstrated that it can improve the result of LSTM for some tasks. Furthermore, we aim to improve our results by employing VisualBert, which is trained on a large multimodal dataset that consists of both text and images. The assumption is that retraining this model on our data would achieve a good result.

REFERENCES

[1] 2008–2021. GROBID. https://github.com/kermitt2/grobid. (2008–2021). swh:1:dir:dab86b296e3c3216e2241968f0d63b68e8209d3c

[2] Philip G. Altbach and Hans de Wit. 2018. Too much academic research is being published. https://www.universityworldnews.com/post.php?story=20180905095203579

[3] 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 (CIKM '17)*. Association for Computing Machinery, New York, NY, USA, 1967–1970. https://doi.org/10.1145/3132847.3133074

[4] Sam Anzaroot and Andrew Mccallum. 2013. A New Dataset for Fine-Grained Citation Field Extraction. *ICML Workshop on Peer Reviewing and Publishing Models*. (2013).

[5] Vidhya Balasubramanian, Sooryanarayan Gobu Doraisamy, and Navaneeth Kumar Kanakarajan. 2016. A Multimodal Approach for Extracting Content Descriptive Metadata from Lecture Videos. *J. Intell. Inf. Syst.* 46, 1 (2016), 121–145. https://doi.org/10.1007/s10844-015-0356-5

[6] Zeyd Boukhers, Nada Beili, Timo Hartmann, Prantik Goswami, and Muhammad Arslan Zafar. 2021. MexPub: Deep Transfer Learning for Metadata Extraction from German Publications. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE.

[7] Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. 2018. Towards Better UD Parsing: Deep Contextualized Word Embeddings, Ensemble, and Treebank Concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, Brussels, Belgium, 55–64. http://www.aclweb.org/anthology/K18-2005

[8] Chien-Chih Chen, Kai-Hsiang Yang, Chuen-Liang Chen, and Jan-Ming Ho. 2012. Bibpro: A citation parser based on sequence alignment. *IEEE Transactions on Knowledge and Data Engineering 24, 2 (2012)* (2012), 236–250.

[9] Pei-Shan Chi. 2014. Which role do non-source items play in the social sciences? A case study in political science in Germany. *Scientometrics* 101, 2 (2014), 1195–1213.

[10] Pei-Shan Chi. 2014. Which role do non-source items play in the social sciences? A case study in political science in Germany. *Scientometrics* 101, 2 (2014), 1195–1213. https://doi.org/10.1007/s11192-014-1433-1

[11] Jason P. C. Chiu and Eric Nichols. 2015. Named Entity Recognition with Bidirectional LSTM-CNNs. *CoRR* abs/1511.08308 (2015). http://arxiv.org/abs/1511.08308

[12] Giovanni Colavizza and Matteo Romanello. 2019. Citation Mining of Humanities Journals: The Progress to Date and the Challenges Ahead. *Journal of European Periodical Studies* 4 (2019), 36–53. https://doi.org/10.21825/jeps.v4i1.10120

[13] Isaac G Councill, C Lee Giles, and Min-Yen Kan. 2008. ParsCit: an Open-source CRF Reference String Parsing Package. *LREC, Vol. 8.* (2008), 661–667.

[14] Min-Yuh Day, Richard Tzong-Han Tsai, Cheng-Lung Sung, Chiu-Chen Hsieh, Cheng-Wei Lee, Shih-Hung Wu, Kun-Pin Wu, Chorng-Shyong Ong, and Wen-Lian Hsu. 2007. Reference metadata extraction using a hierarchical knowledge representation framework. *Decision Support Systems* 43, 1 (2007), 152–167. https://doi.org/10.1016/j.dss.2006.08.006

[15] Hui Han, C.L. Giles, E. Manavoglu, Hongyuan Zha, Zhenyue Zhang, and E.A. Fox. 2003. Automatic document metadata extraction using support vector machines. In *2003 Joint Conference on Digital Libraries, 2003. Proceedings.* 37–48. https://doi.org/10.1109/JCDL.2003.1204842

[16] Hui Han, Eren Manavoglu, Hongyuan Zha, Kostas Tsioutsiouliklis, C. Lee Giles, and Xiangmin Zhang. 2005. Rule-Based Word Clustering for Document Metadata Extraction. In *Proceedings of the 2005 ACM Symposium on Applied Computing* (Santa Fe, New Mexico) *(SAC '05)*. Association for Computing Machinery, New York, NY, USA, 1049–1053. https://doi.org/10.1145/1066677.1066917

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2980–2988. https://doi.org/10.1109/ICCV.2017.322

[18] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *CoRR* abs/1508.01991 (2015). http://arxiv.org/abs/1508.01991

[19] Asanee Kawtrakul and Chaiyakorn Yingsaeree. 2005. A unified framework for automatic metadata extraction from electronic document. In *Proceedings of The International Advanced Digital Library Conference. Nagoya, Japan.*

[20] Huajing Li, Isaac Councill, Wang-Chien Lee, and C Lee Giles. 2006. CiteSeerx: an architecture and web service design for an academic document search engine. In *Proceedings of the 15th international conference on World Wide Web*. 883–884.

[21] 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*. 385–386.

[22] Runtao Liu, Liangcai Gao, Dong An, Zhuoren Jiang, and Zhi Tang. 2018. Automatic Document Metadata Extraction Based on Deep Networks. In *Natural Language Processing and Chinese Computing*, Xuanjing Huang, Jing Jiang, Dongyan Zhao, Yansong Feng, and Yu Hong (Eds.). Springer International Publishing, Cham, 305–317.

[23] Fuchun Peng and Andrew McCallum. 2006. Information Extraction from Research Papers Using Conditional Random Fields. *Inf. Process. Manage.* 42, 4 (2006), 963–979. https://doi.org/10.1016/j.ipm.2005.09.002

[24] Kristie Seymore, Andrew Mccallum, and Ronald Rosenfeld. 1999. Learning Hidden Markov Model Structure for Information Extraction. In *In AAAI 99 Workshop on Machine Learning for Information Extraction*. 37–42.

[25] Christopher G. Stahl, Steven R. Young, Drahomira Herrmannova, Robert M. Patton, and Jack C. Wells. 2018. DeepPDF: A Deep Learning Approach to Extracting Text from PDFs. (2018). https://www.osti.gov/biblio/1460210

[26] Dominika Tkaczyk. 2017. New Methods for Metadata Extraction from Scientific Literature. *CoRR* abs/1710.10201 (2017). arXiv:1710.10201 http://arxiv.org/abs/1710.10201

[27] Dominika Tkaczyk, Paweł Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, and Lukasz Bolikowski. 2015. CERMINE: Automatic Extraction of Structured Metadata from Scientific Literature. *Int. J. Doc. Anal. Recognit.* 18, 4 (2015), 317–335. https://doi.org/10.1007/s10032-015-0249-8

[28] Yi Wu, Edward Y. Chang, Kevin Chen-Chuan Chang, and John R. Smith. 2004. Optimal Multimodal Fusion for Multimedia Data Analysis. In *Proceedings of the 12th Annual ACM International Conference on Multimedia (MULTIMEDIA '04)*. Association for Computing Machinery, New York, NY, USA, 572–579. https://doi.org/10.1145/1027527.1027665

[29] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2016. Aggregated Residual Transformations for Deep Neural Networks. *arXiv preprint arXiv:1611.05431* (2016).

[30] Guixian Xu, Yueting Meng, Xiaoyu Qiu, Ziheng Yu, and Xu Wu. 2019. Sentiment analysis of comment texts based on BiLSTM. *Ieee Access* 7 (2019), 51522–51532.

[31] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1015–1022.