arXiv:2105.05796v1 [cs.CL] 12 May 2021

Kleister: Key Information Extraction Datasets Involving Long Documents with Complex Layouts

Tomasz Stanisławek[1,2], Filip Graliński[1,3], Anna Wróblewska[2],
Dawid Lipiński[1], Agnieszka Kaliska[1,3], Paulina Rosalska[1],
Bartosz Topolski[1], and Przemysław Biecek[2,4]

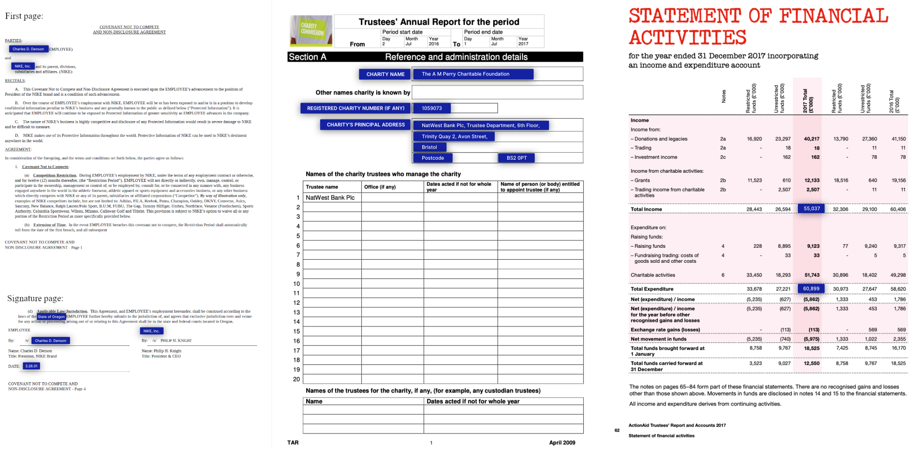
[1] Applica.ai, 15 Zajęcza, Warsaw, 00351 firstname.lastname@applica.ai
[2] Warsaw University of Technology, Koszykowa 75, Warsaw, Poland
firstname.lastname@pw.edu.pl
[3] Adam Mickiewicz University, 1 Wieniawskiego, Poznan, 61712, Poland
firstname.lastname@amu.edu.pl
[4] Samsung R&D Institute Poland, Plac Europejski 1, Warsaw, Poland
firstnameletter.lastname@samsung.com

Abstract. The relevance of the Key Information Extraction (KIE) task is increasingly important in natural language processing problems. But there are still only a few well-defined problems that serve as benchmarks for solutions in this area. To bridge this gap, we introduce two new datasets (*Kleister NDA* and *Kleister Charity*). They involve a mix of scanned and born-digital long formal English-language documents. In these datasets, an NLP system is expected to find or infer various types of entities by employing both textual and structural layout features. The Kleister Charity dataset consists of 2,788 annual financial reports of charity organizations, with 61,643 unique pages and 21,612 entities to extract. The Kleister NDA dataset has 540 Non-disclosure Agreements, with 3,229 unique pages and 2,160 entities to extract. We provide several state-of-the-art baseline systems from the KIE domain (Flair, BERT, RoBERTa, LayoutLM, LAMBERT), which show that our datasets pose a strong challenge to existing models. The best model achieved an 81.77 % and an 83.57 % F1-score on respectively the Kleister NDA and the Kleister Charity datasets. We share the datasets to encourage progress on more in-depth and complex information extraction tasks.

Keywords: Key Information Extraction, Visually Rich Documents, Named Entity Recognition

1 Introduction

The task of Key Information Extraction (KIE) from Visually Rich Documents (VRD) has proved increasingly interesting in the business market with the recent rise of solutions related to Robotic Process Automation (RPA). From a business user's point of view, systems that, fully automatically, gather information about



First page:

COVENANT NOT TO COMPETE
AND NON-DISCLOSURE AGREEMENT

PARTIES:

Charles D. Denson EMPLOYEE)

and

NIKE, Inc. and its parent, divisions, subsidiaries and affiliates. (NIKE):

RECITALS:

A. This Covenant Not to Compete and Non-Disclosure Agreement is executed upon the EMPLOYEE's advancement to the position of President of the NIKE brand and is a condition of such advancement.

B. Over the course of EMPLOYEE's employment with NIKE, EMPLOYEE will be or has been exposed to and/or is in a position to develop confidential information peculiar to NIKE's business and not generally known to the public as defined below ("Protected Information"). It is anticipated that EMPLOYEE will continue to be exposed to Protected Information of greater sensitivity as EMPLOYEE advances in the company.

C. The nature of NIKE's business is highly competitive and disclosure of any Protected Information would result in severe damage to NIKE and be difficult to measure.

D. NIKE makes use of its Protective Information throughout the world. Protective Information of NIKE can be used to NIKE's detriment anywhere in the world.

AGREEMENT:

In consideration of the foregoing, and the terms and conditions set forth below, the parties agree as follows:

1. Covenant Not to Compete.

(a) Competition Restriction. During EMPLOYEE's employment by NIKE, under the terms of any employment contract or otherwise, and for twelve (12) months thereafter, (the "Restriction Period"), EMPLOYEE will not directly or indirectly, own, manage, control, or participate in the ownership, management or control of, or be employed by, consult for, or be connected in any manner with, any business engaged anywhere in the world in the athletic footwear, athletic apparel or sports equipment and accessories business, or any other business which directly competes with NIKE or any of its parent, subsidiaries or affiliated corporations ("Competitor"). *By way of illustration only*, examples of NIKE competitors include, but are not limited to: Adidas, FILA, Reebok, Puma, Champion, Oakley, DKNY, Converse, Asics, Saucony, New Balance, Ralph Lauren/Polo Sport, B.U.M, FUBU, The Gap, Tommy Hilfiger, Umbro, Northface, Venator (Footlockers), Sports Authority, Columbia Sportswear, Wilson, Mizuno, Callaway Golf and Titleist. This provision is subject to NIKE's option to waive all or any portion of the Restriction Period as more specifically provided below.

(b) Extension of Time. In the event EMPLOYEE breaches this covenant not to compete, the Restriction Period shall automatically toll from the date of the first breach, and all subsequent

COVENANT NOT TO COMPETE AND
NON-DISCLOSURE AGREEMENT - Page 1

Signature page:

(d) Applicable Law/Jurisdiction. This Agreement, and EMPLOYEE's employment hereunder, shall be construed according to the laws of the State of Oregon EMPLOYEE further hereby submits to the jurisdiction of, and agrees that exclusive jurisdiction over and venue for any action or proceeding arising out of or relating to this Agreement shall lie in the state and federal courts located in Oregon.

EMPLOYEE	NIKE, Inc.
By: /s/ Charles D. Denson	By: /s/ PHILIP H. KNIGHT
Name: Charles D. Denson Title: President, NIKE Brand	Name: Philip H. Knight Title: President & CEO
DATE: 3.26.01	

COVENANT NOT TO COMPETE AND
NON-DISCLOSURE AGREEMENT - Page 4

CHARITY COMMISSION

Trustees' Annual Report for the period

	Period start date				Period end date		
From	Day 2	Month Jul	Year 2016	To	Day 1	Month Jul	Year 2017

Section A Reference and administration details

CHARITY NAME: The A M Perry Charitable Foundation

Other names charity is known by:

REGISTERED CHARITY NUMBER (IF ANY): 1059073

CHARITY'S PRINCIPAL ADDRESS: NatWest Bank Plc, Trustee Department, 6th Floor,
Trinity Quay 2, Avon Street,
Bristol
Postcode BS2 0PT

Names of the charity trustees who manage the charity

	Trustee name	Office (if any)	Dates acted if not for whole year	Name of person (or body) entitled to appoint trustee (if any)
1	NatWest Bank Plc			
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				

Names of the trustees for the charity, if any, (for example, any custodian trustees)

Name	Dates acted if not for whole year

TAR 1 April 2009

STATEMENT OF FINANCIAL ACTIVITIES

for the year ended 31 December 2017 incorporating
an income and expenditure account

	Notes	Restricted funds (£'000)	Unrestricted funds (£'000)	2017 Total (£'000)	Restricted funds (£'000)	Unrestricted funds (£'000)	2016 Total (£'000)
Income							
Income from:							
– Donations and legacies	2a	16,920	23,297	**40,217**	13,790	27,360	41,150
– Trading	2a	-	18	**18**	-	11	11
– Investment income	2c	-	162	**162**	-	78	78
Income from charitable activities:							
– Grants	2b	11,523	610	**12,133**	18,516	640	19,156
– Trading income from charitable activities	2b	-	2,507	**2,507**	-	11	11
Total Income		28,443	26,594	55,037	32,306	29,100	60,406
Expenditure on:							
Raising funds:							
– Raising funds	4	228	8,895	**9,123**	77	9,240	9,317
– Fundraising trading: costs of goods sold and other costs	4	-	33	**33**	-	5	5
Charitable activities	6	33,450	18,293	**51,743**	30,896	18,402	49,298
Total Expenditure		33,678	27,221	60,899	30,973	27,647	58,620
Net (expenditure) / income		(5,235)	(627)	**(5,862)**	1,333	453	1,786
Net (expenditure) / income for the year before other recognised gains and losses		(5,235)	(627)	**(5,862)**	1,333	453	1,786
Exchange rate gains (losses)		-	(113)	**(113)**	-	569	569
Net movement in funds		(5,235)	(740)	**(5,975)**	1,333	1,022	2,355
Total funds brought forward at 1 January		8,758	9,767	**18,525**	7,425	8,745	16,170
Total funds carried forward at 31 December		3,523	9,027	**12,550**	8,758	9,767	18,525

The notes on pages 65–84 form part of these financial statements. There are no recognised gains and losses other than those shown above. Movements in funds are disclosed in notes 14 and 15 to the financial statements.

All income and expenditure derives from continuing activities.

62 **ActionAid Trustees' Report and Accounts 2017**
Statement of financial activities

Fig. 1. Examples of a real business applications and data for *Kleister* datasets. (Note: The key entities are in blue.)

individuals, their roles, significant dates, addresses and amounts, would be beneficial, whether the information is from invoices or receipts, from company reports or contracts [21,18,13,16,9,12,22]. There is a disparity between what can be delivered with the KIE domain systems on publicly available datasets and what is required by real-world business use. This disparity is still large and makes a robust evaluation difficult. Recently, researchers have started to fill the gap by creating datasets in the KIE domain such as scanned receipts: *SROIE*[5] [18], form understanding [11], NIST Structured Forms Reference Set of Binary Images (*SFRS*)[6] or Visual Question Answering dataset *DocVQA* [15].

This paper describes two new English-language datasets for the Key Information Extraction tasks from a diverse set of texts, long scanned and born-digital documents with complex layouts, that address real-life business problems (Figure 1). The datasets represent various problems arising from the specificity of business documents and associated business conditions, e.g. complex layouts, specific business logic, OCR quality, long documents with multiple pages, noisy training datasets, and normalization. Moreover, we evaluate several systems from the KIE domain on our datasets and analyze KIE tasks' challenges in the business domain. We believe that our datasets will prove a good benchmark for more complex Information Extraction systems.

The main contributions of this study are:

1. *Kleister* – two novel datasets of long documents with complex layouts: 3,328 documents containing 64,872 pages with 23,772 entities to extract (see Section 3);

[5] `https://rrc.cvc.uab.es/?ch=13&com=evaluation&task=3`
[6] `https://www.nist.gov/srd/nist-special-database-2`

2. our method of collecting datasets using a semi-supervised methodology, which reduces the amount of manual work in gathering data; this method has the potential to be reused for similar tasks (see Section 3.1 and 3.2);
3. evaluation over several state-of-the-art Named Entity Recognition (NER) architectures (Flair, BERT, RoBERTa, LayoutLM, LAMBERT) employing our *Pipeline* method (see Section 4.1 and 5);
4. detailed analysis of the data and baseline results related to the Key Information Extraction task carried out by human experts (see Section 3.3 and 5).

The data are available at `https://github.com/applicaai/kleister-nda.git` and `https://github.com/applicaai/kleister-charity.git`.

2 Related Work

Our main reason for preparing a new dataset was to develop a strategy to deal with challenges faced by businesses, which means overcoming such difficulties as complex layout, specific business logic (the way that content is formulated, e.g. tables, lists, titles), OCR quality, document-level extraction and normalization.

2.1 KIE from Visually Rich Documents (publicly available)

A list of KIE-oriented challenges is available at the International Conference on Document Analysis and Recognition ICDAR 2019[7] (cf. Table 1). There is a dataset called SROIE[8] with information extraction from a set of scanned receipts. The authors prepared 1,000 whole scanned receipt images with annotated entities: company name, date, address, and total amount paid (a similar dataset was also created in [18]). Form Understanding in Noisy Scanned Documents is another interesting dataset from ICDAR 2019 (*FUNSD*) [11]. FUNSD aims at extracting and structuring the textual content of forms. However, the authors focus mainly on understanding tables and a limited range of document layouts, rather than on extracting particular entities from the data. The point is, therefore, to indicate a table but not to extract the information it contains.

2.2 KIE from Visually Rich Documents (publicly unavailable)

There are also datasets for the Key Information Extraction task based on invoices [16,17,9,12]. Documents of this kind contain entities like 'Invoice date,' 'Invoice number,' 'Net amount' and 'Vendor Name', extracted using a combination of NLP and Computer Vision techniques. The reason for such a complicated multi-domain process is that spatial information is essential for properly understanding these kinds of documents. However, since they are usually short, the same information is relatively rarely repeated, and therefore there is no need for understanding the more extended context of the document. Nevertheless, those kinds of datasets are the most similar to our use case.

[7] `http://icdar2019.org/competitions-2/`

[8] `https://rrc.cvc.uab.es/?ch=13`

2.3 Information Extraction from one-dimensional documents

The *WikiReading* dataset [8] (and its variant *WikiReading Recycled* [6]) is a large-scale natural language understanding task. Here, the main goal is to predict textual values from the structured knowledge base, Wikidata, by reading the text of the corresponding Wikipedia articles. Some entities can be extracted from the given text directly, but some have to be inferred. Thus, as in our assumptions, the task contains a rich variety of challenging extraction sub-tasks and it is also well-suited for end-to-end models that must cope with longer documents.

Key Information Extraction is different from the Named Entity Recognition task (the *CoNLL 2003* NER challenge [20] being a well-known example). This is because: (1) retrieving spans is not required in KIE; (2) a system is expected to extract specific, actionable data points rather than general types of entities (such as people, organization, locations and "others" for CoNLL 2003).

Dataset name	CoNLL 2003	WikiReading	FUNSD	SROIE	Kleister NDA	Kleister Charity
Source	Reuters news	Wikipedia	forms	receipts	EDGAR	UK Charity Com.
Documents	1,393	4.7M	199	973	540	2,778
Pages	—	—	199	973	3,229	61,643
Entities	35,089	18M	9,743	3,892	2,160	21,612
train docs	946	16.03M	149	626	254	1,729
dev docs	216	1.89M	—	—	83	440
test docs	231	0.95M	50	347	203	609
Input/Output on token level(*)	✓	✗	✓	✓	✗	✗
Long Document(*)	✗	✓	✗	✗	✓	✓
Complex layout(*)	✗	✗	✓	✓	✓	✓
OCR(*)	✗	✗	✓	✓	✗	✓

Table 1. Summary of the existing English datasets and the Kleister sets. (*) For detailed description see Section 3.3.

3 Kleister: New Datasets

We collected datasets of long formal born-digital documents, namely US non-disclosure agreements (Kleister NDA) and a mixture of born-digital and (mostly) scanned annual financial reports of charitable foundations from the UK (Kleister Charity). These two datasets have been gathered in different ways due to their repository structures. Also, they were published on the Internet for different reasons. The crucial difference between them is that the NDA dataset was born-digital, but that the Charity dataset needed to be OCRed. Kleister datasets have a multi-modal input (PDF files and text versions of the documents) and a list of entities to be found.

3.1 NDA Dataset

The NDA Dataset contains Non-disclosure Agreements, also known as Confidentiality Agreements. They are legally binding contracts between two or more parties, where the parties agree not to disclose information covered by the agreement. The NDAs can take on various forms (e.g. contract attachments, emails), but they usually have a similar structure.

Data Collection Method. The NDAs were collected from the Electronic Data Gathering, Analysis and Retrieval system (EDGAR[9]) via Google's search engine. The original files were in an HTML format, but they were transformed into PDF files to keep processing simple and similar to that of other public datasets. Transformation was made using the `puppeteer` library.[10] Then, a list of entities was established (see Table 1).

Annotation Procedure. We annotated the whole dataset in two ways. Its first part, made up of 315 documents, was annotated by three annotators, except that only contexts with some similarity, pre-selected using methods based on semantic similarity (cf. [3]), were taken into account; this was to make the annotation faster and less-labor intensive. The second part, with 195 documents, was annotated entirely by hand. When preparing the dataset, we wanted to determine whether semantic similarity methods could be applied to limit the time it would take to perform annotation procedures; this solution was about 50 % quicker than fully manual annotation. The annotations on all documents were then checked by a super-annotator, which ensured the annotation's excellent quality Cohen's κ (=0.971)[11]. Next, all entities were normalized according to the standards adopted by us, e.g. the `effective date` was standardized according to ISO 8601 i.e. YYYY-MM-DD[12].

Dataset split. The Kleister NDA dataset contains a relatively small document count, so we decided to add more examples into the test split (about 38 %) so as to be more accurate during the evaluation stage (see Table 1 for exact numbers).

3.2 Charity Dataset

The Charity dataset consists of annual financial reports that all charities registered in England and Wales must submit to the Charity Commission. The

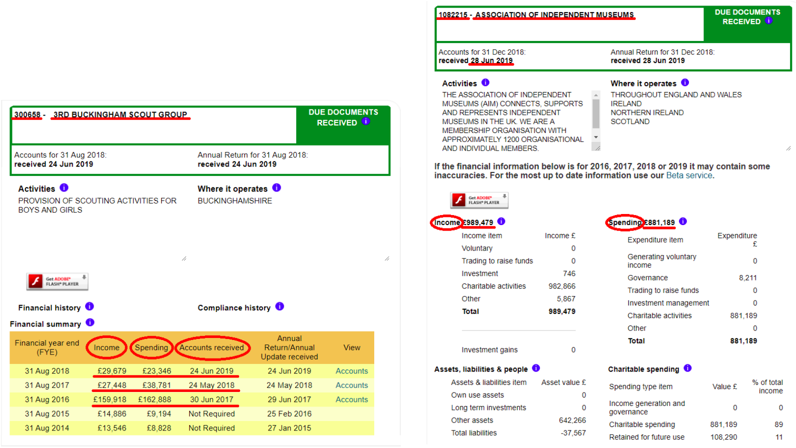
[9] `https://www.sec.gov/edgar.shtml`

[10] `https://github.com/puppeteer/puppeteer`

[11] `https://en.wikipedia.org/wiki/Cohen%27s_kappa`

[12] The normalization standards are described in the public repository with datasets.

Commission subsequently makes them publicly available on its website.[13] There are no strict rules about the format of these charity reports. Some are richly illustrated with photos and charts and financial information constitutes only a small part of the entire report. In contrast, others are a few pages long and only necessary data on revenues and expenses in a given calendar year are given.



300658 - 3RD BUCKINGHAM SCOUT GROUP

DUE DOCUMENTS RECEIVED

Accounts for 31 Aug 2018:
received 24 Jun 2019

Annual Return for 31 Aug 2018:
received 24 Jun 2019

Activities
PROVISION OF SCOUTING ACTIVITIES FOR BOYS AND GIRLS

Where it operates
BUCKINGHAMSHIRE

Financial history

Compliance history

Financial summary

Financial year end (FYE)	Income	Spending	Accounts received	Annual Return/Annual Update received	View
31 Aug 2018	£29,679	£23,346	24 Jun 2019	24 Jun 2019	Accounts
31 Aug 2017	£27,448	£38,781	24 May 2018	24 May 2018	Accounts
31 Aug 2016	£159,918	£162,888	30 Jun 2017	29 Jun 2017	Accounts
31 Aug 2015	£14,886	£9,194	Not Required	25 Feb 2016	
31 Aug 2014	£13,546	£8,828	Not Required	27 Jan 2015	

1082215 - ASSOCIATION OF INDEPENDENT MUSEUMS

DUE DOCUMENTS RECEIVED

Accounts for 31 Dec 2018:
received 28 Jun 2019

Annual Return for 31 Dec 2018:
received 28 Jun 2019

Activities
THE ASSOCIATION OF INDEPENDENT MUSEUMS (AIM) CONNECTS, SUPPORTS AND REPRESENTS INDEPENDENT MUSEUMS IN THE UK. WE ARE A MEMBERSHIP ORGANISATION WITH APPROXIMATELY 1200 ORGANISATIONAL AND INDIVIDUAL MEMBERS.

Where it operates
THROUGHOUT ENGLAND AND WALES
IRELAND
NORTHERN IRELAND
SCOTLAND

If the financial information below is for 2016, 2017, 2018 or 2019 it may contain some inaccuracies. For the most up to date information use our Beta service.

Income £989,479

Income item	Income £
Voluntary	0
Trading to raise funds	0
Investment	746
Charitable activities	982,866
Other	5,867
Total	**989,479**
Investment gains	0

Spending £881,189

Expenditure item	Expenditure £
Generating voluntary income	0
Governance	8,211
Trading to raise funds	0
Investment management	0
Charitable activities	881,189
Other	0
Total	**881,189**

Assets, liabilities & people

Assets & liabilities item	Asset value £
Own use assets	0
Long term investments	0
Other assets	642,266
Total liabilities	-37,567

Charitable spending

Spending type item	Value £	% of total income
Income generation and governance	0	0
Charitable spending	881,189	89
Retained for future use	108,290	11

Fig. 2. Organization's page on the Charity Commission's website (left: organization whose annual income is between 25k and 500k GBP, right: over 500k). Note: Entities are underlined in red and names of entities are circled.

Data Collection Method. The Charity Commission website has a database of all the charity organizations registered in England and Wales. Each of these organizations has a separate sub-page on the Commission's website, and it is easy to find the most important information about them there (see Figure 2). This information only partly overlaps with information in the reports. Some entities such as, say, a list of trustees might not be in the reports. Thus, we decided to extract only those entities which also appear in the form of a brief description on the website.

In the beginning, we downloaded 3,414 reports (as PDF files).[14] During document analysis, it emerged that several reports were written in Welsh. As we are interested only in English, all documents in other languages were identified and removed from the collection. Additionally, documents that contained reports for more than one organization or whose OCR quality was low were deleted. This left us with 2,778 documents.

[13] https://apps.charitycommission.gov.uk/showcharity/registerofcharities/RegisterHomePage.aspx

[14] Organizations with an income below 25,000 GBP a year are required to submit a condoned financial report instead.

Annotation Procedure. There was no need to manually annotate all documents because information about the reporting organizations could be obtained directly from the Charity Commission. Initially, only a random sample of 100 documents were manually checked. Some proved low quality: `charity name` (5 % of errors and 13 % of minor differences), and `charity address` (9 % of errors and 63 % of minor differences). Minor errors are caused by data presentation differences on the page and in the document. For example, the charity's name on the website and in the document could be written with the term *Limited* (shortened to *LTD*) or without it. These minor differences were corrected manually or automatically. In the next step, 366 documents were analyzed manually. Some parts of the charity's address were also problematic. For instance, counties, districts, towns and cities were specified on the website, but not in the documents, or *vice versa*. We split the address data into three separate entities that we considered the most essential: postal code, postal town name and street or road name. The postal code was the critical element of the address, based on the city name and street name[15]. The whole process allowed us to accurately identify entities (see Table 1) and to obtain a good-quality dataset with annotations corresponding to the gold standard.

Dataset split. In the Kleister Charity dataset, we have multiple documents from the same charity organization but from different years. Therefore, we decided to split documents based on charity organization into the train/dev/test sets with, respectively, a 65/15/20 dataset ratio (see Table 1 for exact numbers). The documents from the dev/test split were manually annotated (by two annotators) to ensure high-quality evaluation. Additionally, 100 random documents from the test set were annotated twice to calculate the relevant Cohen's κ coefficient (we achieved excellent quality $\kappa = 0.9$).

Entities	General entity type	Total count	Unique values	(*)Avg. entity count/doc	(*)Avg. token count/entity	Example gold value
NDA dataset (540 documents)						
party	ORG/PER	1,035	912	19.74	1.62	Ajinomoto Althea Inc.
jurisdiction	LOCATION	531	37	1.05	1.21	New York
effective_date	DATE	400	370	1.95	3.10	2005-07-03
term	DURATION	194	22	1.03	2.77	P12M
Charity dataset (2 788 documents)						
post_town	ADDRESS	2,692	501	1.12	1.06	BURY
postcode	ADDRESS	2,717	1,511	1.12	1.99	BL9 ONP
street_line	ADDRESS	2,414	1,353	1.12	2.52	42-47 MINORIES
charity_name	ORG	2,778	1,600	13.80	3.67	Mad Theatre Company
charity_number	NUMBER	2,763	1,514	2.47	1.00	1143209
report_date	DATE	2,776	129	10.58	2.96	2016-09-30
income	AMOUNT	2,741	2,726	1.95	1.01	109370.00
spending	AMOUNT	2,731	2,712	2.03	1.01	90174.00

Table 2. Summary of the entities in the NDA and Charity datasets. (*) Based on manual annotation of text spans.

[15] Postal codes in the UK were aggregated from `www.streetlist.co.uk`

3.3 Statistics and Analysis

The detailed statistics of the Kleister datasets are presented in Table 1 and Table 2. Our datasets covered a broad range of general types of entities; the `party` entity is special since it could be one of the following types: `ORGANIZATION` or `PERSON`. Additionally, some documents may not contain all entities mentioned in the text, for instance in Kleister NDA the `term` entity appears in 36 % of documents. Likewise, some entities may have more than one gold value; for instance in Kleister NDA the `party` entity could have up to 7 gold values for a single document. `Report_date`, `jurisdiction` and `term` have the lowest number of unique values. This suggests that these entities should be simpler than others to extract.

Manual Annotation of Text Spans. To give more detailed statistics we decided to annotate small numbers of documents on text span level. Four annotators annotated 60/55 documents for, respectively, the Kleister Charity and Kleister NDA. In Table 2, we observe that 5 out of 12 entities appear once in a single document. There are also three entities with more than ten counts on average (`party`, `charity_number` and `report_date`). Annotation on the text-span level could prove critical to checking the quality of the training dataset for methods based on a Named Entity Recognition model, something which an *autotagging* mechanism produces (see section 4.1).

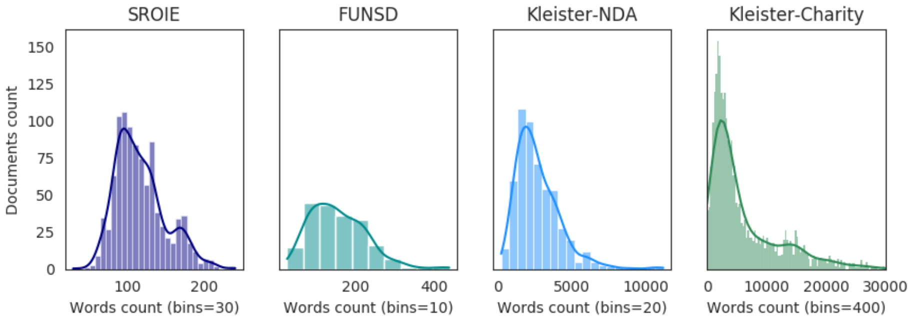


Fig. 3. Distribution of document lengths for Kleister datasets compared to other similar datasets (note that the x-axes ranges are different).

Comparison with Existing Resources. In Table 1, we gathered the most important information about open datasets (which are the most popular ones in the domain) and the Kleister datasets. In particular, we outlined the difference based on the following properties:

- **Input/Output on token level**: it is known which tokens an entity is made up from in the documents. Otherwise, one should: a) create a method for preparing a training dataset for sequence labeling models (subsequently in the publication, we use the term *autotagging* for this sub-task); b) infer or create a canonical form of the final output in order to deal with differences between the target entities provided in the annotations and their variants occurring in the documents (e.g. for `jurisdiction` we must transform a text-level span *NY* into a document-level gold value **New York**).
- **Long Document**: Figure 3 presents differences in document lengths (calculated as a number of OCRed words) in the Kleister datasets compared to other similar datasets. Since entities could appear in documents multiple times in different contexts, we must be able to understand long documents as a whole. This leads, of course, to different architectural decisions [2,4]. For example, the `term` entity in the Kleister NDA dataset tells us about the contract duration. This information is generally found in the *Term* chapter, in the middle part of a document. However, sometimes we could also find a `term` entity at the end of the document, the task is to find out which of the values is correct.
- **Complex Layout**: this requires proper understanding of the complex layout (e.g. interpreting tables and forms as 2D structures), see Fig 1.
- **OCR**: processing of scanned documents in such a way as to deal with possible OCR errors caused by handwriting, pages turned upside down or more general poor scan quality.

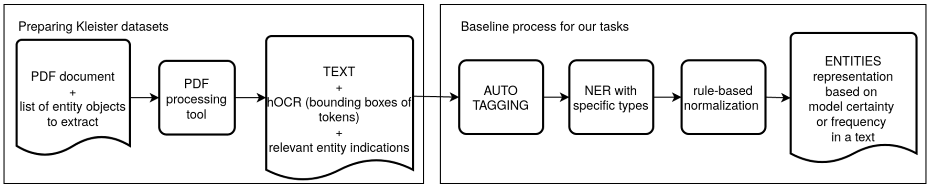

Preparing Kleister datasets
PDF document + list of entity objects to extract
PDF processing tool
TEXT + hOCR (bounding boxes of tokens) + relevant entity indications
Baseline process for our tasks
AUTO TAGGING
NER with specific types
rule-based normalization
ENTITIES representation based on model certainty or frequency in a text

Fig. 4. Our preparation process for Kleister datasets and training baselines. Initially, we gathered PDF documents and expected entities' values. Then, based on textual and layout features, we prepared our pipeline solutions. The pipeline process is illustrated in the second frame and consists of the following stages: autotagging; standard NER; text normalization; and final selection of the values of entities.

4 Experiments

Kleister datasets for the Key Information Extraction task are challenging and hardly any solutions in the current NLP world can solve them. In this experiment, we aim to produce strong baselines with the *Pipeline* approach (Section 4.1) to solve extraction problems. This method's core idea is to select specific parts of the text in a document that denote the objects we search for. The whole process is a chain of techniques with, crucially, a named entity recognition model: once indicated in a document (multiple occurrences are possible), entities are normalized, then all results are aggregated into the one value specified for a given entity type.

4.1 Document Processing Pipeline

Figure 4 presents the whole process, and all the stages are described below (a similar methodology was proposed in [17]).

Autotagging Since we have only document-level annotation in the Kleister datasets, we need to generate a training set for an NER model which takes text span annotation as the input. This stage involves extracting all the fragments that refer to the same or to different entities by using sets of regular expressions combined with a gold-standard value for each general entity type, e.g. date, organization and amount. In particular, when we try to detect a `report_date` entity, we must handle different date formats: 'November 29, 2019', '11/29/19' or '11-29-2019'. This step is performed only for the purpose of training (to get data for training a sequence labeler; it is not applied during the prediction). The quality of this step varies across entity types (see details in Table 3).

Named Entity Recognition. We trained a NER model on the autotagged dataset using one of the state-of-the-art (1) architectures working on plain text such as Flair [1], BERT-base [5], RoBERTa-base [14], or (2) models employing layout features (LayoutLM-base [23] and LAMBERT [7]). Then, at the evaluation stage, we use the NER model to detect all entity occurrences in the text.

Normalization. At this stage objects are normalized to the canonical form, which we have defined in the Kleister datasets. We use almost the same regular expression as during autotagging. For instance, all detected `report_date` occurrences are normalized. So 'November 29, 2019', '11/29/19' and '11-29-2019' are rendered in our standard '2019-11-29' form (ISO 8601).

Aggregation. The NER model might return more than one text span for a given entity, sometimes these are multiple occurrences of the same correct information. Sometimes these represent errors of the NER model. In any case, we need to produce a single output from multiple candidates detected by the NER model. We take a simple approach: all candidates are grouped by the extracted entities' normalized forms and for each group we sum up the scores and finally we return the values with the largest sums.

4.2 Experimental Setup

Due to the Kleister document's length, most currently available models limit input size and so are unable to process the documents in a single pass. Therefore, each document was split into 300-word chunks (for Flair) or 510 BPE tokens (for BERT/RoBERTa/LayoutLM/LAMBERT) with overlapping parts. The results from overlapping parts were aggregated by averaging all the scores obtained for each token in the overlap.

For the Flair-based pipeline, we used implementation from the Flair library [1] in version 0.6.1 with the following parameters: *learning rate* = 0.1, *batch size* = 32, *hidden size* = 256, *epoch* = 30/15 (resp. NDA and Charity), *patience* = 3, *anneal factor* = 0.5, and with a CRF layer on top. For pipeline based on BERT/RoBERTa/LayoutLM, we used the implementation from *transformers* [10] library in version 3.1.0 with the following parameters: *learning rate* = 2e−5, *batch size* = 8, *epoch* = 20, *patience* = 2. For pipeline based on LAMBERT model we used implementation shared by authors of the publication [7] and the same parameters as for the BERT/RoBERTa/LayoutLM models. All experiments were performed with the same settings.

Moreover, in our experiments, we tried different PDF processing tools for text extraction from PDF documents to check the importance of text quality for the final pipeline score:

- **Microsoft Azure Computer Vision API (Azure CV)**[16] – commercial OCR engine, version 3.0.0;
- **pdf2djvu/djvu2hocr**[17]– a free tool for object and text extraction from born-digital PDF files (this is not an OCR engine, hence it could be applied only to Kleister NDA), version 0.9.8;
- **Tesseract**[19] – this is the most popular free OCR engine currently available, we used version 4.1.1-rc1-7-gb36c.[18];
- **Amazon Textract**[19] – commercial OCR engine.

5 Results

Table 3 shows the results for the two Kleister datasets obtained with the Pipeline method for all tested models. The weakest model from our baselines is, in general, BERT, with a slight advantage in Kleister NDA over the Flair model and a large performance drop on Kleister Charity in comparison to others. The best model is LAMBERT, which improved the overall F_1-score with 0.77 and 2.04 for, respectively, NDA and Charity. It is worth noting that for born-digital documents in Kleister NDA this difference is not substantial. This is due to the fact that only for `effective_date` entity does the LAMBERT model have a clear advantage (about 4 points gain of F_1-score) over other baseline models. For Kleister Charity LAMBERT achieves the biggest improvement over sequential models on `income` (+4.03) and `spending` (+5.60) which appears mostly in table-like structures.

[16] `https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/concept-recognizing-text`

[17] `http://jwilk.net/software/pdf2djvu`, `https://github.com/jwilk/ocrodjvu`

[18] run with `--oem 2 -l eng --dpi 300` flags (meaning both new and old OCR engines were used simultaneously, with language and pixel density set to English and 300dpi respectively)

[19] `https://aws.amazon.com/textract/` (API in version from March 1, 2020 was used)

Kleister NDA dataset (pdf2djvu)							
Entity name	**Flair**	**BERT**	**RoBERTa**	**LayoutLM**	**LAMBERT**	Autotagger	Human
effective_date	79.37	80.20	81.50	80.50	**85.27**	79.00	100%
party	70.13	71.60	**80.83**	76.60	78.70	33.15	98%
jurisdiction	93.87	95.00	92.87	94.23	**96.50**	54.10	100%
term	**60.33**	45.73	52.27	47.63	55.03	74.10	95%
ALL	77.83	78.20	81.00	78.47	**81.77**	60.09	97.86%
Kleister Charity dataset (Azure CV)							
post_town	83.07	77.03	77.70	76.57	**83.70**	66.04	98%
postcode	89.57	87.10	88.40	88.53	**90.37**	87.60	100%
street_line	69.10	62.23	72.03	70.92	**74.30**	75.02	96%
charity_name	72.97	75.93	78.03	**79.63**	77.83	67.00	99%
charity_number	96.60	**96.67**	95.37	96.13	95.80	98.60	98%
income	70.67	67.30	69.73	70.40	**74.70**	69.00	97%
report_date	95.93	96.60	96.77	96.40	**96.80**	89.00	100%
spending	68.13	64.43	68.60	68.57	**74.20**	73.00	92%
ALL	81.17	78.33	81.50	81.53	**83.57**	78.16	97.45%

Table 3. The detailed results (average F_1-scores over 3 runs) of our baselines for Kleister challenges (test sets) for the best PDF processing tool. Autotagger F_1-scores were calculated based on results from our regexp mechanism and manual annotation on the text span level (see section 3.3). Human performance is a percentage of annotators agreements for 100 random documents. We used the Base version of the BERT, RoBERTa, LayoutLM and LAMBERT models.

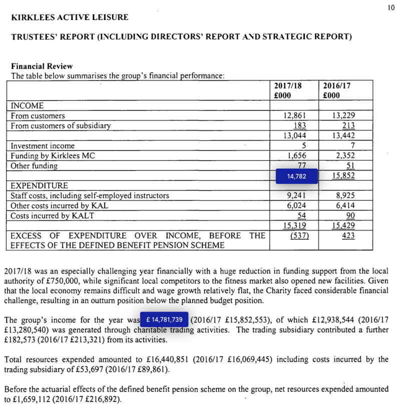
The most challenging problems for all models are entities (`effective_date`, `party`, `term`, `post_town`, `postcode`, `street_line`, `charity_name`, `income`, `spending`) related to the properties described in Section 3.3.

Input/Output on Token Level (Autotagging). As we can observe in Table 3, our autotagging mechanism with information about entity achieves, on the text span level, a performance inferior to almost all our models on the document level. It shows that, despite the fact that the autotagging mechanism is prone to errors, we could train a good quality NER model. Our analysis shows that there are some specific issues related to a regular-expression-based mechanism, e.g. `party` in the Kleister NDA dataset has the lowest score because organization names often occur in the text as an acronym or as a shortened form; for instance for `party` entity text *Emerson Electric Co.* means the same as *Emerson*. This is not easy to capture with a general regexp rule.

Input/Output on token level (normalization). We found that we could not achieve competitive results by using models based only on sequence labeling. For example, for the entities `income` and `spending` in the Kleister Charity dataset, we manually checked that in about 5 % of examples we need to also infer the right scale (thousand, million, etc.) for each monetary value based on the document context (see Figure 5).

Long Documents. It turns out that, for all models, worse results are observed for longer documents, see Figure 6.

Complex Layout. The LAMBERT model has proved the best one, which proved the importance of using models employing not only textual (1D) but also



10

KIRKLEES ACTIVE LEISURE

TRUSTEES' REPORT (INCLUDING DIRECTORS' REPORT AND STRATEGIC REPORT)

Financial Review
The table below summarises the group's financial performance:

	2017/18 £000	2016/17 £000
INCOME		
From customers	12,861	13,229
From customers of subsidiary	183	213
	13,044	13,442
Investment income	5	7
Funding by Kirklees MC	1,656	2,352
Other funding	77	51
	14,782	15,852
EXPENDITURE		
Staff costs, including self-employed instructors	9,241	8,925
Other costs incurred by KAL	6,024	6,414
Costs incurred by KALT	54	90
	15,319	15,429
EXCESS OF EXPENDITURE OVER INCOME, BEFORE THE EFFECTS OF THE DEFINED BENEFIT PENSION SCHEME	(537)	423

2017/18 was an especially challenging year financially with a huge reduction in funding support from the local authority of £750,000, while significant local competitors to the fitness market also opened new facilities. Given that the local economy remains difficult and wage growth relatively flat, the Charity faced considerable financial challenge, resulting in an outturn position below the planned budget position.

The group's income for the year was £ 14,781,739 (2016/17 £15,852,553), of which £12,938,544 (2016/17 £13,280,540) was generated through charitable trading activities. The trading subsidiary contributed a further £182,573 (2016/17 £213,321) from its activities.

Total resources expended amounted to £16,440,851 (2016/17 £16,069,445) including costs incurred by the trading subsidiary of £53,697 (2016/17 £89,861).

Before the actuarial effects of the defined benefit pension scheme on the group, net resources expended amounted to £1,659,112 (2016/17 £216,892).

Fig. 5. Normalization issues for an `income` entity (amount in the table should be multiplied by 1000).

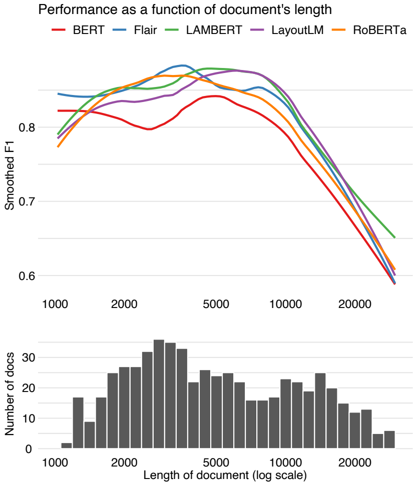


Fig. 6. Relationship between F_1-scores and document length in the Kleister Charity test set for the Azure CV OCR.

layout (2D) features (see Table 3). Additionally, we also observe that the entities appearing in the sequential contexts achieve higher F_1-scores (`charity_number` and `report_date` entities in the Kleister Charity dataset).

OCR. We present the importance of using a PDF processing tool of good quality (see Table 4). With such a tool, we could gain several points in the F_1-score. There are two main conclusions: 1) Commercial OCR engines (Azure CV and Textract) are significantly better than Tesseract for scanned documents (Kleister Charity dataset). This is especially for true for 1D models not trained on Tesseract output (Flair, BERT, RoBERTa); 2) If we have the means to detect born-digital PDF documents, we should process them with a dedicated PDF tool (such as pdf2djvu) instead of using an OCR engine.

6 Conclusions

In this paper, we introduced two new datasets Kleister NDA and Kleister Charity for Key Information Extraction tasks. We set out in detail the process necessary for the preparation of these datasets. Our intention was to show that Kleister datasets will help the NLP community to investigate the effects of document lengths, complex layouts, and OCR quality problems on KIE performance.

We prepared baseline solutions based on text and layout data generated by different PDF processing tools from the datasets. The best model from our baselines achieves 81.77/83.57 F_1-score for, respectively, the Kleister NDA and Charity, which is much lower in comparison to datasets in a similar domain

Kleister NDA dataset (born-digital PDF files)					
PDF tool	**Flair**	**BERT**	**RoBERTa**	**LayoutLM**	**LAMBERT**
Azure CV	$78.03_{\pm 0.12}$	$77.67_{\pm 0.18}$	$79.33_{\pm 0.68}$	$77.43_{\pm 0.29}$	$80.57_{\pm 0.25}$
pdf2djvu	$77.83_{\pm 0.26}$	$78.20_{\pm 0.17}$	$81.00_{\pm 0.05}$	$78.47_{\pm 0.76}$	$\mathbf{81.77}_{\pm \mathbf{0.09}}$
Tesseract	$76.57_{\pm 0.49}$	$76.60_{\pm 0.30}$	$77.81_{\pm 0.97}$	$77.70_{\pm 0.48}$	$81.03_{\pm 0.23}$
Textract	$77.37\ _{\pm 0.08}$	$74.83_{\pm 0.45}$	$79.49_{\pm 0.32}$	$77.40_{\pm 0.40}$	$77.37_{\pm 0.08}$
Kleister Charity dataset (mixture of born-digital and scanned PDF files) (*)					
Azure CV	$81.17_{\pm 0.12}$	$78.33_{\pm 0.08}$	$81.50_{\pm 0.23}$	$81.53_{\pm 0.23}$	$\mathbf{83.57}_{\pm \mathbf{0.29}}$
Tesseract	$72.87_{\pm 0.81}$	$71.37_{\pm 1.25}$	$76.23_{\pm 0.15}$	$77.53_{\pm 0.20}$	$81.50_{\pm 0.07}$
Textract	$78.03_{\pm 0.12}$	$73.30_{\pm 0.43}$	$80.08_{\pm 0.15}$	$80.23_{\pm 0.41}$	$82.97_{\pm 0.21}$

Table 4. F_1-scores for different PDF processing tools and models checked on Kleister challenges test sets over 3 runs with standard deviation. (*) pdf2djvu does not work on scans. We used the Base version of the BERT, RoBERTa, LayoutLM and LAMBERT models.

(e.g. 98.17 [7] for SROIE). This benchmark shows the weakness of the currently available state-of-the-art models for the Key Information Extraction task.

References

1. Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1638–1649. Association for Computational Linguistics, Santa Fe, New Mexico, USA (Aug 2018), `https://www.aclweb.org/anthology/C18-1139`
2. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. ArXiv **abs/2004.05150** (2020)
3. Borchmann, L., Wisniewski, D., Gretkowski, A., Kosmala, I., Jurkiewicz, D., Szalkiewicz, L., Palka, G., Kaczmarek, K., Kaliska, A., Gralinski, F.: Contract discovery: Dataset and a few-shot semantic retrieval challenge with competitive baselines. In: Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020. pp. 4254–4268. Association for Computational Linguistics (2020)
4. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., Salakhutdinov, R.: Transformer-xl: Attentive language models beyond a fixed-length context. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (2019), `https://www.aclweb.org/anthology/P19-1285`
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv **abs/1810.04805** (2018)
6. Dwojak, T., Pietruszka, M., Borchmann, Ł., Chłędowski, J., Graliński, F.: From dataset recycling to multi-property extraction and beyond. In: Proceedings of the 24th Conference on Computational Natural Language Learning. pp. 641–651. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.conll-1.52, `https://www.aclweb.org/anthology/2020.conll-1.52`

7. Garncarek, Ł., Powalski, R., Stanisławek, T., Topolski, B., Halama, P., Turski, M., Graliński, F.: LAMBERT: Layout-Aware (Language) Modeling using BERT for information extraction. ArXiv **abs/2002.08087** (2020)
8. Hewlett, D., Lacoste, A., Jones, L., Polosukhin, I., Fandrianto, A., Han, J., Kelcey, M., Berthelot, D.: WikiReading: A novel large-scale language understanding task over Wikipedia. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1535–1545. Association for Computational Linguistics, Berlin, Germany (2016)
9. Holt, X., Chisholm, A.: Extracting structured data from invoices. In: Proceedings of the Australasian Language Technology Association Workshop 2018. pp. 53–59. Dunedin, New Zealand (Dec 2018), https://www.aclweb.org/anthology/U18-1006
10. Hugging Face: Transformers. https://github.com/huggingface/transformers (2020)
11. Jaume, G., Kemal Ekenel, H., Thiran, J.: FUNSD: A dataset for form understanding in noisy scanned documents. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 2, pp. 1–6 (2019)
12. Katti, A.R., Reisswig, C., Guder, C., Brarda, S., Bickel, S., Höhne, J., Faddoul, J.B.: Chargrid: Towards Understanding 2D Documents. ArXiv **abs/1809.08799** (2018)
13. Liu, X., Gao, F., Zhang, Q., Zhao, H.: Graph convolution for multimodal information extraction from visually rich documents. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (2019), http://dx.doi.org/10.18653/v1/N19-2005
14. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv **abs/1907.11692** (2019)
15. Mathew, M., Karatzas, D., Jawahar, C.V.: DocVQA: A Dataset for VQA on Document Images. ArXiv **abs/2007.00398** (2021)
16. Palm, R.B., Laws, F., Winther, O.: Attend, copy, parse end-to-end information extraction from documents. International Conference on Document Analysis and Recognition (ICDAR) (2019)
17. Palm, R.B., Winther, O., Laws, F.: Cloudscan - a configuration-free invoice analysis system using recurrent neural networks. 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (2017). https://doi.org/10.1109/icdar.2017.74
18. Park, S., Shin, S., Lee, B., Lee, J., Surh, J., Seo, M., Lee, H.: CORD: A Consolidated Receipt Dataset for Post-OCR Parsing. Document Intelligence Workshop at Neural Information Processing Systems (2019)
19. Smith, R.: Tesseract Open Source OCR Engine (2020), https://github.com/tesseract-ocr/tesseract
20. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of the Seventh Conference of the North American Chapter of the Association for Computational Linguistics (2003)
21. Wellmann, C., Stierle, M., Dunzer, S., Matzner, M.: A framework to evaluate the viability of robotic process automation for business process activities. Business Process Management: Blockchain and Robotic Process Automation Forum (2020)
22. Wróblewska, A., Stanisławek, T., Prus-Zajączkowski, B., Garncarek, Ł.: Robotic process automation of unstructured data with machine learning. In: Position Papers of the 2018 Federated Conference on Computer Science and Information

Systems, FedCSIS 2018, Poznań, Poland, September 9-12, 2018. pp. 9–16 (2018). https://doi.org/10.15439/2018F373

23. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: LayoutLM: Pre-training of text and layout for document image understanding. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2020). https://doi.org/10.1145/3394486.3403172