

# End-to-End Document Classification and Key Information Extraction using Assignment Optimization

Ciaran Cooney      Joana Cavadas      Liam Madigan      Bradley Savage  
Rachel Heyburn      Mairead O’Cuinn

Aflac Northern Ireland

{ccooney, jcavadas, lmadigan, bsavage,  
rheyburn, mocuinn}@aflac.com

## Abstract

We propose end-to-end document classification and key information extraction (KIE) for automating document processing in forms. Through accurate document classification we harness known information from templates to enhance KIE from forms. We use text and layout encoding with a cosine similarity measure to classify visually-similar documents. We then demonstrate a novel application of mixed integer programming by using assignment optimization to extract key information from documents. Our approach is validated on an in-house dataset of noisy scanned forms. The best performing document classification approach achieved 0.97 f1 score. A mean f1 score of 0.94 for the KIE task suggests there is significant potential in applying optimization techniques. Abation results show that the method relies on document preprocessing techniques to mitigate Type II errors and achieve optimal performance.

## 1 Introduction

For many organizations, the ubiquity of personal computers and smart devices has led to the digitization of processes that previously required human involvement. In particular, paper documents are being digitized more frequently to enable electronic processing. Among those being digitized, documents such as invoices or insurance forms are common in daily workflows but often require tedious and costly manual processing or suffer from brittle automation systems (Majumder et al., 2020). Automating workflows through rule-based systems or machine learning techniques can reduce the requirement for employees to engage in time-consuming data-entry and archiving tasks while cutting costs for employers (Audebert et al., 2019; Mandivarapu et al., 2021). Document classification and key information extraction (KIE) are important tasks in automated document processing, and the two are not mutually exclusive. Classification of document

images is often an important first stage in enterprise document processing upon which more granular downstream processing rests (Zhang and Zhang, 2020; Appalaraju et al., 2021; Audebert et al., 2019; Bakkali et al., 2020; Kumar et al., 2014; Harley et al., 2015; Noce et al., 2016). KIE is one of those downstream tasks with the goal of automating retrieval of key information from documents (Majumder et al., 2020; Gao et al., 2021a,b; Schuster et al., 2013; Chiticariu et al., 2013; Palm et al., 2019; Garncarek et al., 2020).

Document image classification has been studied for decades (Peng et al., 2003; Sarkar, 2010; Shin et al., 2001; Taylor et al., 1995). Early approaches considered functional landmarks and visual features of the spatial layout as discriminating factors between documents (Taylor et al., 1995; Hu et al., 1999; Shin et al., 2001). This strategy treats document images holistically and has been successful in classifying broad categories of documents (Taylor et al., 1995; Shin et al., 2001; Shin and Doermann, 2006), but is less discriminating when applied to structurally- and visually-similar form-like documents (Harley et al., 2015). For these, templating techniques have sometimes been employed (Peng et al., 2001, 2003; Sarkar, 2010). Templating can be defined as finding the highest similarity between any document and a predefined set of document templates (Peng et al., 2003).

However, following computer vision-inspired developments in deep learning resulting from AlexNet (Krizhevsky et al., 2012) and ImageNet (Deng et al., 2009), many approaches essentially treated document classification as a deep learning image classification problem (Afzal et al., 2015; Kumar et al., 2012, 2014; Kang et al., 2014; Tensmeyer and Martinez, 2017; Kölsch et al., 2017). Document images enable extraction of features such as graphics, typeface and colour, not possible with text-only approaches. Convolutional Neural Networks (CNN) were the primary driver

of this development, with one of the initial applications being a 4-layer CNN used to classify tax forms and the Small Tobacco Dataset (Kang et al., 2014). Other works used ImageNet pre-training as a starting-point for CNN-based classifiers (Harley et al., 2015; Afzal et al., 2015). However, Tensmeyer and Martinez (2017) argue that domain differences between natural images and documents limit the efficacy of this approach. In scenarios where broad document categories with distinct visual styles need to be classified, general features of each group positively assist classification (e.g., scientific papers differ greatly from magazine articles). However, within-group documents (e.g., forms from within a company - the category being investigated in this work) can exhibit high visual similarity and therefore require more granularity for accurate classification.

Increased granularity has been provided by a recent upsurge in multimodal approaches to document classification. Methods include combining textual representations with visual features (Noce et al., 2016; Bakkali et al., 2020, 2022; Audebert et al., 2019), combining textual representations with positional embeddings (Xu et al., 2020; Wang et al., 2022; Li et al., 2021a) or combinations of all three (Appalaraju et al., 2021; Xu et al., 2021; Huang et al., 2022; Gu et al., 2021; Li et al., 2021b; Powalski et al., 2021). Audebert et al. (2019) demonstrated that their multimodal approach improves upon two unimodal baselines, while a dual-stream document classification technique combined word embeddings and visual features, with late fusion used to learn joint representations of documents (Bakkali et al., 2020).

LayoutLM, a large pre-trained model for document analysis, models interactions between text and layout information, with and without visual representations (Xu et al., 2020). At the time of publication, this approach achieved SOTA results in document image classification when combining all three modalities. Inspired by this, StructuralLM uses cell level 2d positional embeddings and token embeddings and a novel cell position training objective to achieve 96.08% on RVL-CDIP (Li et al., 2021a). Another text and layout approach, LiLT, uses bi-directional attention to model cross-modal interactions to achieve similar performance (95.68%) (Wang et al., 2022).

Other general document analysis approaches, such as LayoutLMv2 and LayoutLMv3 have

adapted the way visual features are incorporated in multimodal approaches, including using masked image modelling and word-patch alignment training objectives to ensure cross modal alignment (Xu et al., 2021; Huang et al., 2022). Despite having proven that visual features such as text font and visual layout are useful for extracting general differences between documents, it is not clear that they perform as well when differentiating between very similar documents, a common problem in industry applications. Additionally, these large pre-trained models require significant computing resources and large training datasets which are not always available. We address some of these issues by using a classification approach that relies on document templates rather than large-scale training.

Often downstream of document classification, KIE, including value retrieval, involves the extraction of key information from structured documents such as forms or invoices. Extracted information can be used for many tasks including customer enrollment and insurance claim adjudication. Several early approaches to KIE relied on templates constructed to enable cross-referencing of new documents against a bank of existing templates (Rusinol et al., 2013; Chiticariu et al., 2013). Rusinol et al. (2013) developed user-generated training samples which they used to build a document model based on structural templates. Other approaches relied on pre-registered templates within the system to perform KIE (Chiticariu et al., 2013; Schuster et al., 2013). However, Chiticariu et al. (2013) showed a disconnect between industry and academia w.r.t. the utility of rules-based approaches to IE. Although industry applications often require some form of rules-based intervention in production systems, template and rule-based methods are constrained to specific layouts and may not generalize well to unseen documents.

In the era of deep learning it is common to formulate KIE as a sequence-labelling task (Huang et al., 2015; Lample et al., 2016). However, this approach does not handle complex spatial relationships and is not ideal for highly structured documents (Hwang et al., 2021). Due to this there have been several developments in deep learning approaches that have gone beyond the sequence-labelling formulation. Palm et al. (2019) proposed a CNN approach to KIE that rejected the need for word-level labels, and therefore is useful in real-world scenarios where labelled data is not always available.

Another novel approach, Doc2Dict is a T5 transformer trained on database records to produce a document-to-data-structure model [Townsend et al. \(2021\)](#). [Zhang and Zhang \(2020\)](#) proposed an end-to-end text reading and KIE method that extracts entity values through a multimodal fusion of text and visual features, and [Majumder et al. \(2020\)](#) presented a field-value pairing framework that utilizes knowledge about the data-types of fields to select candidate entities. In the latter, a set of candidates is identified as possibly corresponding to a field in the target schema. A neural network is used to learn representations of each candidate based on neighbouring words, before selecting a correct value. This approach is useful in scenarios with high volumes of unseen documents.

As the structure of forms is often more sophisticated than a simple linear ordering of tokens, relative token position, paragraph spacing, and question-answer relationships all contain information that can be leveraged for KIE tasks ([Garncarek et al., 2020](#)). Chargrid ([Katti et al., 2018](#)), CUTIE ([Zhao et al., 2019](#)) and BERTgrid ([Denk and Reisswig, 2019](#)) were among the first models to integrate 2D representations of word tokens alongside text for KIE, and each outperformed their respective baselines. Utilizing the 2D positions of text to assist KIE has helped architectures such as LAMBERT ([Garncarek et al., 2020](#)) and the LayoutLM family ([Xu et al., 2020, 2021](#); [Huang et al., 2022](#)) achieve SOTA performance while exhibiting less sensitivity to serialization. Whereas LayoutLM uses tokens, layout and visual information, LAMBERT relies only on tokens and bounding boxes.

This work introduces an end-to-end document classification and KIE pipeline based on a templating approach that does not require any model training. We eschew the recent trend towards deep learning focused approaches to document classification and KIE while retaining the important consideration of 2D document structural layout. We demonstrate that different text encoding strategies work remarkably well for document classification when computing cosine similarity between a document and a set of document templates. A novel assignment optimization technique for KIE is presented which assigns values in a form to a corresponding template key based on global geometric positions and specified constraints. This approach is insensitive to serialization and word tokenization and does not require large training data while being easily

updated. We report a series of processing steps that are required to make assignment optimization feasible for noisy scanned documents, and test these with ablation experiments. Finally, we detail some limitations of our approach and suggest future directions for refining the technique further.

This work makes the following contributions:

- We present a novel assignment optimization approach to key information extraction.
- We present a granular document classification strategy by finding the cosine similarity between a vectorized document and a matrix of document templates.
- We demonstrate the importance of document rotation and entity scaling processing steps in maximising the potential of our optimization approach.

## 2 Related Work

### 2.1 Document Classification

Features can be extracted from document images to represent either text content or visual properties. Image templates have often been favoured in industrial document analysis ([Sarkar, 2010](#)). [Sarkar \(2010\)](#) use image anchor templates for document classification and propose a method for learning templates from few training examples. Combining both text and visual content is a popular approach to document classification. [Noce et al. \(2016\)](#) reported performance improvement when supplementing visual features with text, especially in cases where different classes share visual characteristics. StructuralLM ([Li et al., 2021a](#)) and LiLT ([Wang et al., 2022](#)) are deep learning approaches that demonstrated the value of combining text and layout information for document classification. With the proliferation of text encoding techniques such as Word2Vec ([Mikolov et al., 2013](#)) and ELMO ([Sarzynska-Wawer et al., 2021](#)) several studies have used off-the-shelf algorithms to generate representations of document images for classification ([Bakkali et al., 2020](#); [Audebert et al., 2019](#)).

### 2.2 Key Information Extraction

KIE has received focused attention, with several early approaches relying on templates constructed to enable cross-referencing of new documents against a bank of existing templates ([Rusinol et al.,](#)

2013; Chiticariu et al., 2013; Schuster et al., 2013). Recently, deep learning approaches have dominated the literature. Palm et al. (2019) present a deep neural network that bypasses the need for word-level training labels, with the assumption that the same structured information must be extracted from each document. Another approach seeks to predict target values from arbitrary queries (Gao et al., 2021a). The technique utilises a novel pre-training strategy which makes it more flexible for learning local geometric relations between words. Gao et al. (2021b) also propose a novel framework for using unlabelled documents for training with known form types. Attempting to overcome expensive annotation efforts, the approach uses a bootstrapped training process to develop a rule-based approach for getting pseudo-labels from unlabelled forms. Pseudo-labels are then used for supervised learning. Related to this approach for reducing annotation effort, findings that bounding boxes alone can be effective in VrDU tasks (Cooney et al., 2023), and work formulating KIE as a constrained optimization problem through partial graph matching (Yao et al., 2021), we propose an assignment optimization approach to KIE.

### 2.3 Assignment Optimization

An assignment problem has a number of *agents* and a number of *tasks*. Agents are assigned to perform tasks, with a cost depending on agent-task assignment. Gong et al. (2021) optimize passenger-route assignment by formulating it as a nonlinear mixed integer optimization model. Other problems such as resource allocation can be formulated as assignment problems to assign a number of tasks (e.g. jobs), to a number of workers (agents), to maximise or minimise some utility function (Lee et al., 2018). In our formulation, agents are template value-positions corresponding to a specific key. Tasks are form entities. The cost we seek to minimize is the Euclidian distance between template value-positions and form entities.

## 3 Dataset

Datasets used in document classification studies typically consist of broad categories of documents (e.g. email, letter, scientific report) (Harley et al., 2015; Kumar et al., 2014). Our data consists of 395 scanned document images, each corresponding to one of six categories of health insurance claim forms and exhibiting a degree of structural

Form	N documents	N key-values
aicf_pg1	169	2484
aicf_pg2	144	1390
hicf_pg1	30	422
hicf_pg2	24	352
aicf_v1	14	217
pvcf	14	279

Table 1: Dataset statistics for each form type.

homogeneity common within organizations (Figure 1). Each form contains a set of *keys* adjacent to an associated text-box or white-space into which information (*values*) can be entered. Each key in a form asks a claimant to enter a specific piece of information (e.g. first name, last name, policy number), but not all value spaces have been filled-in. Forms in the dataset contain both printed and handwritten responses to key requests for information. The scanned documents are of fixed dimensions ( $1700 \times 2200$  pixels) and exhibit significant rotation variance and noise. The number of key-value pairs per document class ranges from 217 to 2484. Dataset statistics are reported in Table 1.

We use Amazon Textract<sup>1</sup> Optical Character Recognition (OCR) engine to extract text and bounding boxes from the scanned documents and the six document templates.

### 3.1 Consolidating OCR Output

Previous studies have noted the negative impact that OCR mistakes can have on KIE tasks (Audebert et al., 2019; Palm et al., 2019). Due to the noisy nature of many of our scanned documents, OCR output can be degraded in terms of identifying characters that form part of an entity string (Figure 2 - top). To deal with this, we implement a simple yet effective method for consolidating character strings that are likely to form part of a single entity. First, we iterate through the set of strings identified by OCR from top-left to bottom-right of the document. For each source string (i.e., the string we may wish to append to), we search potential candidate strings to determine whether they are vertically aligned with the source. Here, we use a threshold of  $\pm 15$  pixels for both top and bottom of the bounding boxes. If bounding boxes are inline, we then compute the distance along the horizontal axis between the trailing character in the source string and the leading character in the candidate string. If this distance is below a specified threshold, the character strings are consolidated, and their

<sup>1</sup><https://aws.amazon.com/textract/>



OCR. We implemented three different methods for vectorizing the document text: two versions of the Universal Sentence Encoder (USE) (Cer et al., 2018) and Term Frequency Inverse Document Frequency<sup>2</sup> (TF-IDF).

The USE approaches are designed to be general purpose text embedding models. Both USE models receive input English strings and return a fixed 512 element vector representation of the string as a sentence embedding, which here represents the entire document. The first USE method is a transformer-based encoding model which generates text embeddings by using attention to compute context aware representations of words (Vaswani et al., 2017). The second is a deep averaging network which computes average input embeddings for words and bi-grams before being passed through a feedforward deep neural network to obtain a 512 element sentence embedding.

TF-IDF computes the relative frequency of words in one document compared to the inverse frequency of that word in the entire corpus. In this case the document corpus consists of the combined text from the six forms (Section 3). The TF-IDF algorithm is fitted to the data to learn its vocabulary and inverse document frequency. Each of the six forms is then transformed into a document-term matrix using the fitted vectorizer. The fitted vectorizer is retained to transform new unseen forms.

Forms often exhibit similar structure and when taken from a single domain can often contain duplication of text (e.g., institutional logos, requests for similar information). Figure 3 depicts the correlation between USE vector representations of the template documents, indicating the difficulty in distinguishing between forms using this approach. In particular, forms *aicf\_pg1* and *hicf\_pg1* (0.943) and forms *aicf\_pg2* and *hicf\_pg2* (0.934) are highly correlated. Here we supplement our text embeddings by also generating layout embeddings for each form. These embeddings represent the 2D positions of the text and are generated using the pretrained Layoutlmv2 model (Xu et al., 2021).

## 4.2 Document Classification using Cosine Similarity

Document class is predicted by measuring the cosine similarity between a new document and a set of document templates. A template matrix  $u$  is

<sup>2</sup>[https://scikit-learn.org/stable/modules/generated/sklearn.feature\\_extraction.text.TfidfVectorizer.html](https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html)

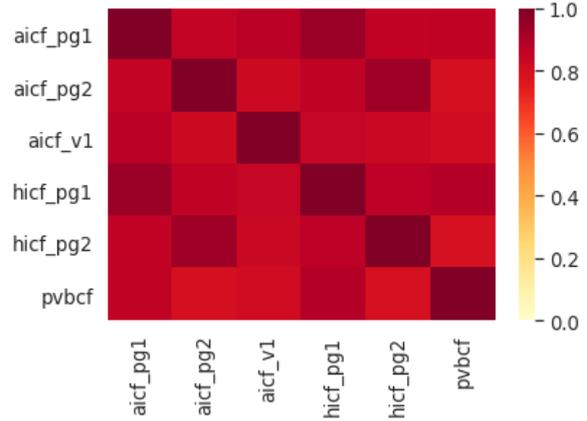


Figure 3: Correlation between embedded text representations of document templates is high.

constructed as a  $n \times m$  matrix where  $n$  is the number of templates in the template bank and  $m$  is the length of the vector representation of a single document. Each new form representation to be measured against the template matrix is represented as a  $1 \times m$  vector  $v$ . Cosine distance between vectors  $u$  and  $v$  is formulated as:

$$s(u, v) = 1 - \frac{u \cdot v}{\|u\|_2 \|v\|_2} \quad (1)$$

where,  $\|*\|_2$  is the 2-norm of argument  $*$ , and  $u \cdot v$  is the dot product of  $u$  and  $v$ . This returns a  $n \times 1$  matrix containing similarity between the candidate form  $v$  and  $n$  templates in matrix  $u$ . We then take the *argmax* of the similarity matrix to obtain a document classification.

## 4.3 Problem Formulation for KIE

Document classification is followed by KIE, where templates are constructed for each of the six document classes in our dataset. The goal of KIE is to identify entity values in a document that are associated with a specific key. Given a document template:

$$DT = \left\{ \left( kw^{(i)}, vx_{min}^{(i)}, vy_{min}^{(i)}, vx_{max}^{(i)}, vy_{max}^{(i)} \right) \mid i \in \{1, \dots, n\} \right\}$$

consisting of  $n$  key words  $kw^{(i)}$  and a bounding box associated with their corresponding value positions  $vx_{min}^{(i)}, vy_{min}^{(i)}, vx_{max}^{(i)}, vy_{max}^{(i)}$ , and a document:

$$D = \left\{ \left( w^{(j)}, x_{min}^{(j)}, y_{min}^{(j)}, x_{max}^{(j)}, y_{max}^{(j)} \right) \mid j \in \{1, \dots, m\} \right\}$$

consisting of  $m$  words  $w^{(j)}$  and their associated bounding box  $x_{min}^{(j)}, y_{min}^{(j)}, x_{max}^{(j)}, y_{max}^{(j)}$ , the objective is to correctly assign  $n$  document bounding boxes to each of the template bounding boxes. In doing so, we assign associated words (values) to the set of key words in the template, thus producing key-value pairs.

Each KIE template consists of a set of key-value pairs. Template keys are text (e.g. 'Name', 'D.O.B.'), and values are a bounding box corresponding to a location on the page where information can be entered. With our method of applying assignment optimization to the KIE task, each bounding box is represented by a single coordinate point. In initial experiments, the bounding box centroid was selected as the 2D coordinate to represent the position of each entity. However, this approach proved unstable during optimization as differences in length between unfilled template values and unseen form values routinely resulted in unexpected mis-assignments. Our reported results are based on selection of the upper-left point of each bounding box representing an entity position, as used by [Katti et al. \(2018\)](#). The reason for this is that while the length of form value entries vary significantly and therefore skew the centroid position, the starting position of an entity value is relatively fixed, providing a more stable representation.

#### 4.4 Scaling Document Entities

A virtue of using accurate document classification to select templates for KIE is that prior information on the specific form being processed can be used to enhance optimization. An issue with scanned documents is significant variation in scale, rotation, and aspect ratio in comparison with the original template versions of these forms ([Ahmad et al., 2021](#)). Here, we use prior knowledge from form templates to scale unseen form entities to approximate dimensions of the template (Figure 4).

As a first step, we compute coordinates for twenty rectangular segments of equal area. This segmentation is required because scanned documents are not always warped linearly e.g., entities at the top of a page can be offset by a different factor to those at the bottom. Then, within each

segment we search for text strings extracted by OCR. Next, we check whether a word within a given segment is present among the set of keys in our entity template. For this, we implement fuzzy matching with a minimum confidence level of 0.9 to account for errors in OCR outputs. To avoid mis-scaling due to duplicate terms within a document, we apply a maximum distance threshold.

Within each segment, Manhattan Distance is measured between each text string and a matching template keyword:

$$d(x, y) = |x_1 - x_2| + |y_1 - y_2| \quad (2)$$

where,  $(x_1, y_1)$  and  $(x_2, y_2)$  are 2d coordinates of the top-left of template and new form entity bounding boxes, respectively. Mean scalar values for each axis are calculated for each segment independently, resulting in distinct vertical and horizontal scalars for scaling each segment. Using these scalars, we re-scale the unseen form entities to approximate positions in the template (Figure 4).

#### 4.5 Assignment Optimization

To assign entities from a new form to keys in a template, we developed a binary optimization model. This approach eliminates any requirement for data-hungry neural network training. The optimization objective is to minimise the distance between entities in a new form and spaces in a form template where values are entered i.e., when a text string in a form has been entered into one of the template value positions, we expect to see a close-to-zero distance between the two bounding boxes. Of course, text in a form that does not correspond to some filled-in value is expected to have a relatively high distance to our template bounding boxes.

Let  $i \in T_k$  and  $j \in F_l$  be a value-position within a template  $T_k \in \mathcal{T}$  and an entity within the form  $F_l \in \mathcal{F}$ , respectively. The standard euclidean distance  $D_{ij}$  between  $i$  and  $j$  is defined as (3), with  $\mathcal{T}_i^k$  and  $\mathcal{F}_j^l$  being the coordinates of value-position and entity bounding boxes, respectively.

$$D_{ij} = \mathcal{T}_i - \mathcal{F}_j, i \in T, j \in F \quad (3)$$

The optimization model used to assign form entities to template value-positions comprises the following features:

- $\mathcal{T}$ : form template.
- $\mathcal{T}_i = (t_i^x, t_i^y)$ : coordinates of value-position  $i$  within template (top-left corner).

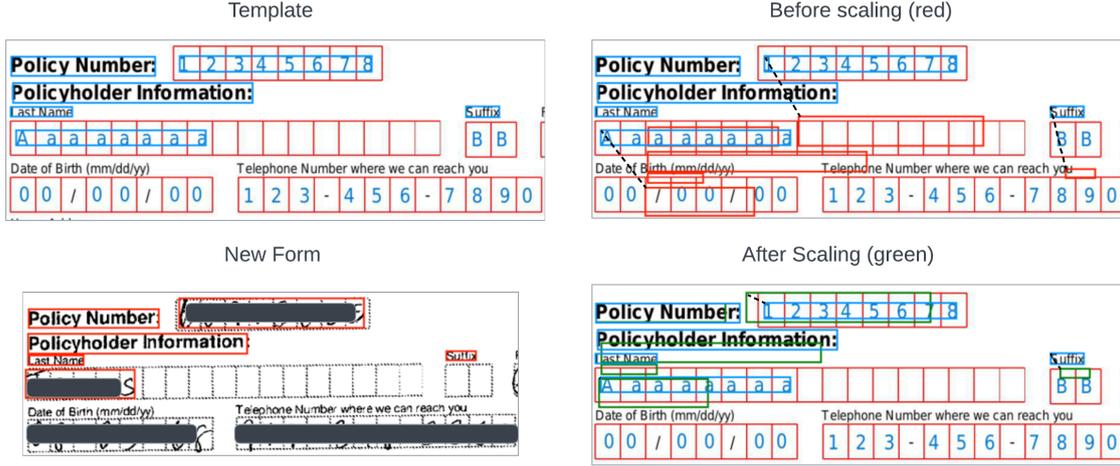


Figure 4: Entities in a scanned document form are scaled to more closely approximate the assigned template. Red bounding boxes indicate scanned document bounding box positions pre-scaling. Green boxes indicate positions post-scaling.

- $\mathcal{F}$ : form.
- $\mathcal{F}_j = (f_j^x, f_j^y)$ : coordinates of entity  $j$  within form (top-left corner).
- $\mathcal{D}_{ij}$ : distance between value-position  $i$  from template value-position  $T_i$  and entity  $j$  from form entity  $F_j$ .
- $\mathcal{M}_{ij}$ : set of form entities matching template keys.

$$\min \sum_{i \in \mathcal{T}} \sum_{j \in \mathcal{F}} \mathcal{D}_{ij} \cdot x_{ij} \quad (4)$$

$$\text{s.to } \sum_{j \in \mathcal{F}} x_{ij} = 1, \quad i \in \mathcal{T} \quad (5)$$

$$\sum_{i \in \mathcal{T}} x_{ij} \leq 1, \quad j \in \mathcal{F} \quad (6)$$

$$x_{ij} = M_{ij}, \quad i \in \mathcal{T}, j \in \mathcal{F}, \quad (7)$$

$$x_{ij} \in \{0, 1\} \text{ for all } i \text{ in } \mathcal{T} \text{ and all } j \text{ in } \mathcal{F} \quad (8)$$

where (4) tries to minimize global distance between template value bounding boxes and new form bounding boxes by assigning a maximum of one form bounding box to one template bounding box. The final output is a set of pairs  $x_{ij}$ , where  $i$  is a template bounding box and  $j$  is a form bounding box, such that  $x_{ij} = 1$ .

Figure 5 depicts the implementation of assignment optimization for the KIE task. Bounding-boxes for template key-value positions (e.g. 'Policy

		Form Entities								
		"Policy Number"	"0123456789"	"Jane"	"Date of Birth"	"First Name"	"Last Name"	"Doe"	"00/00/0000"	"Zip Code"
Template Keys	Policy Number	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	First Name	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	Last Name	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>				
	Date of Birth	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
	Zip Code	<input type="checkbox"/>	<input checked="" type="checkbox"/>							

Figure 5: Assignment optimization for form entities to template keys. Rows are template positions, columns are new form positions. Red squares are constraints indicating a column can't be assigned to a row. Green squares indicate assignment.

Number', 'Last Name') are designated rows, while form entity (e.g. 'Doe', '0123456789') bounding boxes are designated columns. Our optimization approach assigns entities to keys by minimizing overall distance between form entities and template value-positions, with constraints. Green squares indicate that a form entity has been assigned to a template key. Red indicates a constraint on assigning certain entities to certain keys. White squares indicate that an assignment is possible but has not been made for this solution.

## 5 Results

The transformer-based USE using text and layout encodings achieved a weighted average f1 score of

0.97 for the document classification task. In total, 383 out of 395 forms were assigned to the correct template label. The deep random network and TF-IDF approaches also performed well with f1 scores of 0.95 and 0.94, respectively. Form *hicf\_pg2* accounts for the majority of misclassified instances, conforming with similarity scores between templates (Figure 3).

Precision, recall, and f1 scores are used to evaluate the performance of our assignment optimization approach to KIE (Table 2). With a mean f1 score of 0.941, our method exhibits strong performance for the KIE task despite the noise and rotation variance present in our dataset. Mean precision and recall scores of 0.954 and 0.928 suggest that the current implementation of the approach is slightly more likely to misclassify entities in the form than assign values to spaces that were actually unfilled. This means we can have a high degree of confidence that values are assigned to the correct keys, but less confidence that all entities have been assigned.

A striking result from Table 2 is the difference in precision (0.962) and recall (0.872) for *hicf\_pg2*. The reason for this relatively poor recall score appears to be the density of text surrounding the key-value pairs in that particular form. The close proximity of extraneous text to values to be extracted, and the imperfect document scaling technique combine to produce a relatively large number of misassignments. Another notable result is the performance of our method at extracting information from the *aicf\_v1* form. The proposition that KIE algorithms perform more or less effectively depending on the type of document they face is not new (Zhao et al., 2019; Cheng et al., 2022).

Results are comparable with other KIE techniques. On the SROIE dataset, LayoutLM, Struct-Text, LayoutLMv2, TILT, LAMBERT achieved mean KIE f1 scores ranging from 0.952 - 0.982 (Xu et al., 2020; Li et al., 2021c; Xu et al., 2021; Powalski et al., 2021; Garncarek et al., 2020), and TRIE++ recently reported 0.984 (Cheng et al., 2022). Although f1 scores presented here are slightly inferior, our approach offers the benefit of very specific key labelling that can be important for industrial applications of IE.

## 5.1 Ablation Study

To evaluate the impact of preprocessing steps on the overall performance of our KIE method we performed an ablation study. In separate experi-

Document	Precision	Recall	F1
aicf_pg1	0.962 (2312/2404)	0.931 (2312/2484)	0.946
aicf_pg2	0.950 (1304/1372)	0.938 (1304/1390)	0.944
aicf_v1	0.887 (182/205)	0.839 (182/217)	0.862
hicf_pg1	0.954 (393/412)	0.931 (393/422)	0.942
hicf_pg2	0.962 (307/319)	0.872 (307/352)	0.915
pvbcbf	0.939 (275/293)	0.986 (275/279)	0.962
<b>Mean</b>	0.954	0.928	0.941

Table 2: Precision (TP/(TP+FP)), Recall (TP/(TP+FN)) and F1 scores for each document type.

Document	Precision	Recall	F1
aicf_pg1	0.921	0.747	0.825
aicf_pg2	0.952	0.791	0.864
aicf_v1	0.894	0.734	0.806
hicf_pg1	0.925	0.874	0.899
hicf_pg2	0.948	0.786	0.859
pvbcbf	0.857	0.768	0.810
<b>Mean</b>	0.916	0.783	0.844

Table 3: Precision, Recall and F1 when realignment is not applied.

ments, we remove document image alignment (Section 3.2) and entity scaling (Section 4.4) from the pipeline. Results in Table 3 demonstrate an almost 10% drop in mean F1 score when documents are not aligned with template images. A similar decrease in performance is observed when entity scaling is removed from the pipeline (Table 4).

The degraded performance is largely due to an increase in Type II errors (false negatives), as witnessed by a sharp decline in recall scores. This vindicates the use of these preprocessing steps to augment our assignment optimization algorithm that would otherwise struggle with noisy and skewed scanned documents. Without rotation and scaling our algorithm fails to classify approximately 20% of key entities we should extract from our forms.

## 6 Limitations

A limitation of this work is that it does not fully consider how to integrate extraction of information from tables using our optimization approach. Several recent studies have reported entity extraction

Document	Precision	Recall	F1
aicf_pg1	0.923	0.745	0.825
aicf_pg2	0.958	0.824	0.886
aicf_v1	0.918	0.774	0.840
hicf_pg1	0.922	0.839	0.879
hicf_pg2	0.983	0.841	0.906
pvbcbf	0.909	0.860	0.883
<b>Mean</b>	0.936	0.788	0.856

Table 4: Precision, Recall and F1 when document scaling is not applied.

from tables (Paliwal et al., 2019; Yang et al., 2022; Nazir et al., 2021; Agarwal et al., 2021), suggesting feasible approaches to this task. For example, Katti et al. (2018) and Denk and Reisswig (2019) tackle tabular data extraction by predicting the coordinates of the table rows bounding boxes to identify the invoiced products. Another limitation is that skew detection and document dewarping are not currently a fully automated process in our pipeline. Methods for skew detection and correction, such as those reported in Ahmad et al. (2021) could be incorporated in future work to aid processing. Our method has been tested on a limited number of form classes, and currently it does appear to be weaker when faced with densely-populated forms and documents. This is unsurprising, but methods for dealing with this are required to make this method generalise to a more diverse set of forms.

Another potential limitation of the system is that the Transformer-based USE is more expensive than the deep averaging network in terms of compute time (Cer et al., 2018). The deep averaging network compute time is linear with length of string, but classification accuracy was not as strong as the transformer approach. Document classification based on similarity assigns a class even when a document is not one of the classes. In a production system, we would consider an 'other' category for such cases.

## 7 Future Work

Results indicate that our optimization approach to KIE is not entirely robust to all document structures. It will be important to experiment with techniques to improve the generalizability of this method to a more diverse range of documents. Reducing sensitivity to closely located text will help reduce some of the Type II errors. Additionally, it is possible that setting distance constraints between templates and new forms along two dimensions may enhance our assignment operation.

Another area of future work will seek to improve the input to our optimizer. OCR engines can have difficulty dealing with obscure fonts, diacritics, or other relatively uncommon artefacts (Audebert et al., 2019), and we know that the method proposed here relies on strong OCR performance to achieve reliable results. One of the areas to improve our approach is to enhance the fidelity of the initial text recognition phase. This can take the form of using domain-specific knowledge to

make more accurate corrections and using an OCR engine optimized for handwriting recognition.

Finally, we think there is scope within this approach to add semantic information to the assignment of values to keys. Currently, the approach relies solely on positions to make assignments. However, we have seen from many multimodal approaches to document understanding that text is highly informative (Xu et al., 2020, 2021; Garncarek et al., 2020; Cooney et al., 2023; Li et al., 2021c,a). LAMBERT uses word tokens alongside bounding boxes for KIE (Garncarek et al., 2020), and Majumder et al. (2020) use the data-type of fields to enhance value retrieval. Further work is required to investigate ways in which text can be integrated into our KIE approach.

## 8 Conclusions

In this paper, we present an end-to-end document classification and KIE technique that uses a novel application of assignment optimization to extract key information from insurance claim forms. The system is end-to-end in that accurate document classification is required to select a specific template to be applied downstream to the KIE task. We experiment with several encoding methods for documents, and use cosine similarity to measure the distance between a form and a bank of templates. Key values are extracted from forms and assigned to a corresponding value position in a template using an optimization algorithm. The noisy scanned documents used to validate the approach require substantial preprocessing through realignment and scaling.

A mean f1 score of 0.94 indicates that this is a promising new approach to KIE from structured forms. An ablation study indicated preprocessing stages are essential to optimize performance of this approach. Our analysis of results obtained from different documents has suggested potential areas where this approach can be enhanced with further development. The approach is particularly suited to industrial applications in which large volumes of identical forms with different information require extraction of key information.

## References

Muhammad Zeshan Afzal, Samuele Capobianco, Muhammad Imran Malik, Simone Marinai, Thomas M Breuel, Andreas Dengel, and Marcus

- Liwicki. 2015. Deepdocclassifier: Document classification with deep convolutional neural network. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 1111–1115. IEEE.
- Madhav Agarwal, Ajoy Mondal, and CV Jawahar. 2021. Cdec-net: Composite deformable cascade network for table detection in document images. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9491–9498. IEEE.
- Riaz Ahmad, Saeeda Naz, and Imran Razzak. 2021. Efficient skew detection and correction in scanned document images through clustering of probabilistic hough transforms. *Pattern Recognition Letters*, 152:93–99.
- Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 993–1003.
- Nicolas Audebert, Catherine Herold, Kuider Slimani, and Cédric Vidal. 2019. Multimodal deep networks for text and image-based document classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 427–443. Springer.
- Souhail Bakkali, Zuheng Ming, Mickaël Coustaty, and Marçal Rusiñol. 2020. Visual and textual deep feature fusion for document image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 562–563.
- Souhail Bakkali, Zuheng Ming, Mickael Coustaty, Marçal Rusiñol, and Oriol Ramos Terrades. 2022. Vl-dcoc: Vision-language contrastive pre-training model for cross-modal document classification. *arXiv preprint arXiv:2205.12029*.
- Gary Bradski and Adrian Kaehler. 2000. Opencv. *Dr. Dobb's journal of software tools*, 3:120.
- Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.
- Zhanzhan Cheng, Peng Zhang, Can Li, Qiao Liang, Yunlu Xu, Pengfei Li, Shiliang Pu, Yi Niu, and Fei Wu. 2022. Trie++: Towards end-to-end information extraction from visually rich documents. *arXiv preprint arXiv:2207.06744*.
- Laura Chiticariu, Yunyao Li, and Frederick Reiss. 2013. Rule-based information extraction is dead! long live rule-based information extraction systems! In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 827–832.
- Ciaran Cooney, Rachel Heyburn, Liam Madigan, Mairead O’Cuinn, Chloe Thompson, and Joana Cavadas. 2023. Unimodal and multimodal representation training for relation extraction. In *Artificial Intelligence and Cognitive Science: 30th Irish Conference, AICS 2022, Munster, Ireland, December 8–9, 2022, Revised Selected Papers*, pages 450–461. Springer.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.
- Timo I Denk and Christian Reisswig. 2019. Bertgrid: Contextualized embedding for 2d document representation and understanding. In *Workshop on Document Intelligence at NeurIPS 2019*.
- Martin A Fischler and Robert C Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395.
- Mingfei Gao, Zeyuan Chen, Nikhil Naik, Kazuma Hashimoto, Caiming Xiong, and Ran Xu. 2021a. Field extraction from forms with unlabeled data. *arXiv preprint arXiv:2110.04282*.
- Mingfei Gao, Le Xue, Chetan Ramaiah, Chen Xing, Ran Xu, and Caiming Xiong. 2021b. Value retrieval with arbitrary queries for form-like documents. *arXiv preprint arXiv:2112.07820*.
- Lukasz Garncarek, Rafal Powalski, Tomasz Stanislawek, Bartosz Topolski, Piotr Halama, and Filip Gralinski. 2020. Lambert: Layout-aware language modeling using bert for information extraction. *arxiv abs/2002.08087 (2020)*.
- Manlin Gong, Yucong Hu, Zhiwei Chen, and Xiaopeng Li. 2021. Transfer-based customized modular bus system design with passenger-route assignment optimization. *Transportation Research Part E: Logistics and Transportation Review*, 153:102422.
- Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. 2021. Unidoc: Unified pretraining framework for document understanding. *Advances in Neural Information Processing Systems*, 34:39–50.
- Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995. IEEE.
- Jiaying Hu, Ramanujan Kashi, and Gordon Wilfong. 1999. Document image layout comparison and classification. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR’99 (Cat. No. PR00318)*, pages 285–288. IEEE.

- Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. *arXiv preprint arXiv:2204.08387*.
- Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.
- Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang, and Minjoon Seo. 2021. Spatial dependency parsing for semi-structured document information extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 330–343.
- Le Kang, Jayant Kumar, Peng Ye, Yi Li, and David Doermann. 2014. Convolutional neural networks for document image classification. In *2014 22nd International Conference on Pattern Recognition*, pages 3168–3172. IEEE.
- Anoop Raveendra Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2d documents. *arXiv preprint arXiv:1809.08799*.
- Andreas Kölsch, Muhammad Zeshan Afzal, Markus Ebbecke, and Marcus Liwicki. 2017. Real-time document image classification using deep cnn and extreme learning machines. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 1318–1323. IEEE.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.
- Jayant Kumar, Peng Ye, and David Doermann. 2012. Learning document structure for retrieval and classification. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 1558–1561. IEEE.
- Jayant Kumar, Peng Ye, and David Doermann. 2014. Structural similarity for document image classification and retrieval. *Pattern Recognition Letters*, 43:119–126.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.
- Mengyuan Lee, Yuanhao Xiong, Guanding Yu, and Geoffrey Ye Li. 2018. Deep neural networks for linear sum assignment problems. *IEEE Wireless Communications Letters*, 7(6):962–965.
- Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2021a. Structurallm: Structural pre-training for form understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6309–6318.
- Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. 2021b. Selfdoc: Self-supervised document representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5652–5660.
- Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. 2021c. Structext: Structured text understanding with multi-modal transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1912–1920.
- Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, James Bradley Wendt, Qi Zhao, and Marc Najork. 2020. Representation learning for information extraction from form-like documents. In *proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 6495–6504.
- Jaya Krishna Mandivarapu, Eric Bunch, Qian You, and Glenn Fung. 2021. Efficient document image classification using region-based graph neural network. *arXiv preprint arXiv:2106.13802*.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Surabhi Narayan and Sahana D Gowda. 2017. Document image registration for imposed layer extraction. *ICTACT Journal on Image & Video Processing*, 7(3).
- Danish Nazir, Khurram Azeem Hashmi, Alain Pagani, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. 2021. Hybridtabnet: Towards better table detection in scanned document images. *Applied Sciences*, 11(18):8396.
- Lucia Noce, Ignazio Gallo, Alessandro Zamberletti, and Alessandro Calefati. 2016. Embedded textual content for document image classification with convolutional neural networks. In *Proceedings of the 2016 ACM Symposium on Document Engineering*, pages 165–173.
- Shubham Singh Paliwal, D Vishwanath, Rohit Rahul, Monika Sharma, and Lovekesh Vig. 2019. Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 128–133. IEEE.

- Rasmus Berg Palm, Florian Laws, and Ole Winther. 2019. Attend, copy, parse end-to-end information extraction from documents. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 329–336. IEEE.
- Hanchuan Peng, Fuhui Long, and Zheru Chi. 2003. Document image recognition based on template matching of component block projections. *IEEE Transactions ON pattern Analysis and machine Intelligence*, 25(9):1188–1192.
- Hanchuan Peng, Fuhui Long, Zheru Chi, and Wan-Chi Siu. 2001. Document image template matching based on component block list. *Pattern Recognition Letters*, 22(9):1033–1042.
- Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. 2021. Going full-tilt boogie on document understanding with text-image-layout transformer. In *International Conference on Document Analysis and Recognition*, pages 732–747. Springer.
- Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee.
- Marçal Rusinol, Tayeb Benkhelfallah, and Vincent Poulain dAndecy. 2013. Field extraction from administrative documents by incremental structural templates. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1100–1104. IEEE.
- Prateek Sarkar. 2010. Learning image anchor templates for document classification and data extraction. In *2010 20th International Conference on Pattern Recognition*, pages 3428–3431. IEEE.
- Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michał Jarkiewicz, and Łukasz Okruszek. 2021. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135.
- Daniel Schuster, Klemens Muthmann, Daniel Esser, Alexander Schill, Michael Berger, Christoph Weidling, Kamil Aliyev, and Andreas Hofmeier. 2013. Intellix—end-user trained information extraction for document archiving. In *2013 12th International Conference on Document Analysis and Recognition*, pages 101–105. IEEE.
- Christian Shin, David Doermann, and Azriel Rosenfeld. 2001. Classification of document pages using structure-based features. *International Journal on Document Analysis and Recognition*, 3(4):232–247.
- Christian K Shin and David S Doermann. 2006. Document image retrieval based on layout structural similarity. In *IPCV*, pages 606–612.
- Suzanne Liebowitz Taylor, Mark Lipshutz, and Roslyn Weidner Nilson. 1995. Classification and functional decomposition of business documents. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 2, pages 563–566. IEEE.
- Chris Tensmeyer and Tony Martinez. 2017. Analysis of convolutional neural networks for document image classification. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 388–393. IEEE.
- Benjamin Townsend, Eamon Ito-Fisher, Lily Zhang, and Madison May. 2021. Doc2dict: Information extraction as text generation. *arXiv preprint arXiv:2105.07510*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Jiapeng Wang, Lianwen Jin, and Kai Ding. 2022. Lilt: A simple yet effective language-independent layout transformer for structured document understanding. *arXiv preprint arXiv:2202.13669*.
- Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2021. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591.
- Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200.
- Jingfeng Yang, Aditya Gupta, Shyam Upadhyay, Luheng He, Rahul Goel, and Shachi Paul. 2022. Tableformer: Robust transformer modeling for table-text encoding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 528–537.
- Minghong Yao, Zhiguang Liu, Liangwei Wang, Houqiang Li, and Liansheng Zhuang. 2021. One-shot key information extraction from document with deep partial graph matching. *arXiv preprint arXiv:2109.13967*.
- Haopeng Zhang and Jiawei Zhang. 2020. Text graph transformer for document classification. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Xiaohui Zhao, Endi Niu, Zhuo Wu, and Xiaoguang Wang. 2019. Cutie: Learning to understand documents with convolutional universal text information extractor. *arXiv preprint arXiv:1903.12363*.