Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)
Distributed under a CC BY-NC-SA 4.0 license.

Predicting Semantic Labels of Text Regions in Heterogeneous Document Images

Somtochukwu Enendu
University of Twente
Enschede, The Netherlands
senendu5@yahoo.com

Johannes Scholtes
ZyLAB
Amsterdam, The Netherlands
Johannes.Scholtes@zylab.com

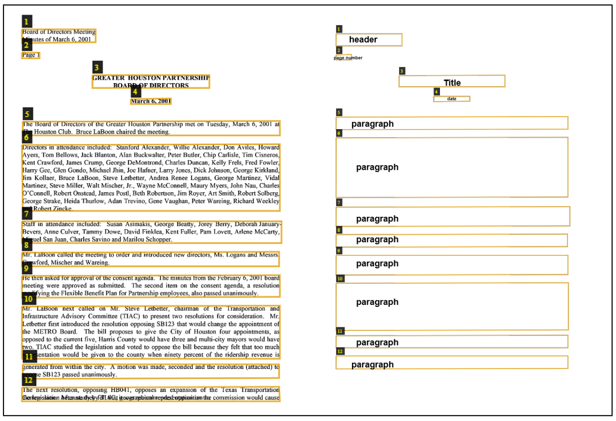
Jeroen Smeets
ZyLAB
Amsterdam, The Netherlands
Jeroen.Smeets@zylab.com

Djoerd Hiemstra
Radboud University Nijmegen
Nijmegen, The Netherlands
Djoerd.Hiemstra@ru.nl

Mariet Theune
University of Twente
Enschede, The Netherlands
m.theune@utwente.nl

Abstract

This paper describes the use of sequence labeling methods in predicting the semantic labels of extracted text regions of heterogeneous electronic documents, by utilizing features related to each semantic label. In this study, we construct a novel dataset consisting of real world documents from multiple domains. We test the performance of the methods on the dataset and offer a novel investigation into the influence of textual features on performance across multiple domains. The results of the experiments show that the neural network method slightly outperforms the Conditional Random Field method with limited training data available. Regarding generalizability, our experiments show that the inclusion of textual features aids performance improvements.



Board of Directors Meeting
Minutes of March 6, 2001

Page 1

GREATER HOUSTON PARTNERSHIP
BOARD OF DIRECTORS

March 6, 2001

The Board of Directors of the Greater Houston Partnership met on Tuesday, March 6, 2001 at The Houston Club. Bruce LaBoon chaired the meeting.

Directors in attendance included: Stanford Alexander, Willie Alexander, Don Aviles, Howard Ayers, Tom Bellows, Jack Blanton, Alan Buckwalter, Peter Butler, Chip Carlisle, Tim Cisneros, Kent Crawford, James Crump, George DeMontrond, Charles Duncan, Kelly Frels, Fred Fowler, Harry Gee, Glen Gondo, Michael Jhin, Joe Hafner, Larry Jones, Dick Johnson, George Kirkland, Jim Kollaer, Bruce LaBoon, Steve Letbetter, Andrea Renee Logans, George Martinez, Vidal Martinez, Steve Miller, Walt Mischer, Jr., Wayne McConnell, Maury Myers, John Nau, Charles O'Connell, Robert Onstead, James Postl, Beth Robertson, Jim Royer, Art Smith, Robert Solberg, George Strake, Heida Thurlow, Adan Trevino, Gene Vaughan, Peter Wareing, Richard Weekley and Robert Zincke.

Staff in attendance included: Susan Asimakis, George Beatty, Jorey Berry, Deborah January-Bevers, Anne Culver, Tammy Dowe, David Finklea, Kent Fuller, Pam Lovett, Arlene McCarty, Miguel San Juan, Charles Savino and Marilou Schopper.

Mr. LaBoon called the meeting to order and introduced new directors, Ms. Logans and Messrs. Crawford, Mischer and Wareing.

He then asked for approval of the consent agenda. The minutes from the February 6, 2001 board meeting were approved as submitted. The second item on the consent agenda, a resolution modifying the Flexible Benefit Plan for Partnership employees, also passed unanimously.

Mr. LaBoon next called on Mr. Steve Letbetter, chairman of the Transportation and Infrastructure Advisory Committee (TIAC) to present two resolutions for consideration. Mr. Letbetter first introduced the resolution opposing SB123 that would change the appointment of the METRO Board. The bill proposes to give the City of Houston four appointments, as opposed to the current five, Harris County would have three and multi-city mayors would have two. TIAC studied the legislation and voted to oppose the bill because they felt that too much representation would be given to the county when ninety percent of the ridership revenue is

generated from within the city. A motion was made, seconded and the resolution (attached) to oppose SB123 passed unanimously.

The next resolution, opposing HB041, opposes an expansion of the Texas Transportation [illegible] commission would cause

Figure 1: Example of a segmented document and its corresponding labels

1 Introduction

On a daily basis, legal departments of corporations produce many electronic documents for documentation of cases, investigative reporting, internal communication etc. Whenever these corporations are involved in litigation or investigations as part of regulatory requests, the need arises to collect and review these documents and share their contents with third parties. As document data sets increase, the corporations turn to e-discovery technology to facilitate the process of collecting, reviewing and sharing. E-discovery technology helps to automatically analyze the documents by using text mining and other text-related analytics to discover relevant information. However, these text mining techniques for automatic document analysis only work optimally when the roles of different text sections in a document are known. For example, by recognizing tables, headers and footers, we can apply different extraction and analysis techniques than on normal paragraphs, and expect better results.

For safety reasons however, electronic documents in the legal domain are mostly transformed into images (e.g. jpg, tiff) so the corporation or firm can have control of what they share with other parties. Electronic documents usually contain hidden information (information that can't be seen when the document is viewed) and these pieces of information could contain hidden details they don't want to disclose to the receiving party. On the other hand, transforming the documents to images creates another problem as it makes it more difficult to automatically identify the specific role of the document areas. Hence, to provide automatic tools to determine the function of textual regions derived from document images, we need to do document image understanding.

The primary goal in document image understanding is to (1) identify regions of interest in a document image (page segmentation) and (2) recognize the role of each region (semantic structure labeling). Many related studies treat these two tasks as separate sequential tasks. However, they are also often handled as one unified task. In this work, we specifically address the second step in the understanding of document images: the task of semantic structure labeling. The goal of this task is to label a sequence of physically segmented regions in a document image with semantic labels such as header, paragraph, footer, caption, etc. (see Figure 1). We treat the task as a sequence labeling problem, which involves assigning a categorical label to each member of a sequence of observations i.e. a sequence of document segments in our scenario. Though the work of document image understanding covers various types of document images, our work focuses on electronic and digital-born documents composed primarily of single-column layouts. Typical examples of such electronic documents which can be converted to images are PDF, Word, Powerpoint, E-mails, etc.

Even though extracting the semantic information from a document is a task that is easily done by a human, it is still an open and challenging problem for computers due to the inherent complexity of documents (Rangoni et al., 2012), especially when the set of documents in focus are diverse in layout and format. Similar works on semantic labeling such as (Tao et al., 2013) and (Shetty et al., 2007) are usually very specific to a document format or a set of related document types and problematic when we try to generalize to other document types. There is still a need for robust methods, capable of dealing with a broad spectrum of layouts found in digital-born documents (Clausner et al., 2011).

Our work addresses this gap in research by comparing sequential labeling methods for the semantic labeling task, and considering heterogeneous document images. Homogenous formats and lack of fine-grained semantic labels relevant for real world documents, are some limitations of previous document image datasets. To address these issues, we annotated a new dataset containing documents from an infamous legal case - the Enron Corporation scandal investigation. We also compare the performance of the following sequence labeling methods on the annotated dataset: (i) A feature-based Conditional Random Field (CRF) (ii) A recurrent neural network with a Bidirectional Long Short-Term Memory (LSTM) architecture.

Our methods perform fine-grained recognition on text regions and include identification of tables. Furthermore, we check the influence of textual related features on the generalizability of our methods to a different domain. Luong et al. (2010) and Yang et al. (2017) prove that the performance of methods improves when text information in a region is considered for semantic labeling. We extend this by checking its influence across a different document domain.

Our main contributions are summarized as follows:

- We compare two sequential labeling methods to address document semantic structure labeling. Unlike previous works, we consider heterogeneous document formats and identify both fine-grained semantic-based classes and tables.

- We offer a novel investigation into the influence of text-related features on the performance of our methods across a different document domain.

- We provide an evaluation dataset for the task of semantic labeling on digital-born documents.[1]

In section 3, we present our evaluation dataset. We then provide a detailed description of our system architecture in section 4. Section 5 is a breakdown of the sequence labeling methods performed for the task. We show the results of our experiments in section 6 and conclude on our work in section 7.

2 Related Work

Previous works on document image understanding (Chen and Blostein, 2007; Marinai, 2008; Kamola et al., 2015) divide the task into two parts: a physical decomposition or segmentation of document images into regions (page segmentation) and a logical/semantic understanding of these regions (semantic structure labeling). Though the focus of our work is on semantic labeling, we also present a high-level discussion on existing page segmentation techniques.

[1] The dataset will be made available upon request.

2.1 Page Segmentation

Page segmentation techniques involve identifying segments enclosing homogeneous content regions, such as text, table, figure or graphic in a document page or image. These techniques fall into three categories: *bottom-up*, *top-down* and *hybrid* approaches. Bottom-up approaches (Kise et al., 1998; Adnan and Ricky, 2011) begin by grouping pixels of interest and merging them into larger blocks or connected components, which are then clustered into words, lines or blocks of text. However, such approaches are expensive from a computational point of view. Top-down approaches (Antonacopoulos, 1998; Gatos et al., 1999) recursively segment large regions in a document into smaller sub regions. Both approaches however, are limited by their inability to successfully segment complex and irregular document layouts. Hybrid methods, such as proposed in Pavlidis and Zhou (1992) combine both top-down and bottom-up techniques. With recent advances in deep neural networks, neural based models have become state-of-the-art for segmentation. Siegel et al. (2018) utilized a neural network to extract figures and captions from scientic documents. Yang et al. (2017) proposed a unified convolutional model to classify pixels in a document based on their visual appearance and underlying text content.

2.2 Semantic Structure Labeling

Our work focuses on the second aspect of document image understanding. Semantic labeling couples semantic meaning to a physical region or zone of a document after it has been segmented. Two types of approaches have been considered in the literature to handle this task: the *model-driven approach* and the *data-driven approach* (Mao et al., 2003). Early work in semantic structure labeling focused on the model driven approach. Models made up of rules, or trees, or grammars contained all the information that was used to transform a physical structure into a logical or semantic one. Rule based systems (Kim et al., 2000), though fast and human-understandable proved to be poorly flexible and unable to handle irregular cases and varying layouts.

Recent studies have considered the data-driven approach using supervised learning methods as an alternative to avoid the inflexibility and rigidity of manually built rule systems and mechanisms. These data-driven approaches make use of raw physical data to analyze the document and no knowledge or predefined rules are given. Various document image datasets have been created for this purpose including images in the document space of electronic documents, scanned documents, magazines, newspapers etc. (Todoran et al., 2005; Antonacopoulos et al., 2009) but they are usually confined to a single domain or class. Chen et al. (2007) define a document space as the set of documents that a classifier is expected to handle. The labeled training and test samples are all drawn from this document space. Our dataset includes heterogeneous formats of electronic documents such as Microsoft Office files, PDF and email files which cover multiple domains like business letters, articles, memos, forms, reports, invoices etc. that significantly vary in layout and content.

Most existing supervised learning methods for semantic labeling use CRF and deep neural network approaches. Tao et al. (2013) built a CRF model as a graph structure to label fragments in a document. Shetty et al. (2007) used CRFs utilizing contextual information to automatically label extracted segments from a document. Yang et al. (2017) and Stahl et al. (2018) used visual cues and deep learning methods to analyze documents. In this study, we treat the semantic structure labeling task as a sequential labeling problem where a document image is modeled as a sequence of regions. The motivation for this is to model spatial dependencies and possible transitions between the different regions. Shetty et al. (2007) model spatial inter-dependencies between sequential segments in documents. Luong et al. (2010) also treat their semantic labeling task as an instance of the sequential labeling problem. CRFs and recurrent neural networks are popular sequential learning methods for this type of modeling. We offer a comparison of these state-of-the-art methods for semantic labeling across heterogeneous document formats in this study.

Luong et al. (2010) report in their work that adding textual information to a CRF model for semantic labeling improves its performance. We build on this work by also checking the influence of textual information on the performance of our methods across different document domains.

3 Datasets

This section describes the construction of our evaluation dataset for the task of semantic labeling

Dataset	SemLab	PRIMA
Document images	400	478
Document space	Office docs, PDF & Email	Magazine
Label categories	13	9

Table 1: Overview of the datasets used in this study.

which we call SemLab (SemLab coined from Semantic Labeling). The documents we used were gathered from the Enron Corpus.This corpus is a large database of approximately 600,000 emails generated by 158 employees of the Enron Corporation and acquired by the Federal Energy Regulatory Commission, a United States federal agency, during its investigation after the company's collapse.

To compare the performance of the sequence labeling methods across different domains, we used the PRIMA dataset of Antonacopoulos et al. (2009). Table 1 contains an overview of both datasets.

3.1 Dataset Creation

We select documents for our dataset from the email folder of the then CEO of Enron corporation. Of all the employees in the corporation, he received the most emails. The documents comprise of sent and received email messages in the folder as well as document attachments. For attached documents, we consider four formats of documents: Word, PDF, Excel and Powerpoint documents, and ignore other file formats in the folder. This selection of different document formats meets the *variety* characteristic of an ideal dataset as described in Antonacopoulos et al. (2006) because several classes of document pages are represented. In total, we select 100 email messages and 406 unique documents from the CEO's email folder. With each document containing different pages, the full set we collected from the email folder contained 2,447 document pages.

After selection of the electronic documents, we converted them to TIFF images since document images are the focus of our work. The SemLab evaluation dataset is a random selection of 400 documents from the 2,447 document images, containing a total of 2,869 regions and their ground truth representation in CSV format (see section 3.3).

3.2 Document Semantic Labels

We attempt to identify 13 labels in a document: *paragraph*, *page header*, *caption*, *section heading*, *footer*, *page number*, *table*, *list item*, *title*, *email header*, *email body text*, *email signature* and *email footer*. Our choice of labels is specific to regions in a document that contain text. Hence we didn't consider regions in a document that are devoid of text e.g. figure, image, graphic etc.

3.3 Annotation Process

Apart from the document images part of our dataset, we created the geometric hierarchical structure of each image (in CSV format) as ground truth for the dataset. We achieved this as follows: For each region, the corresponding bounding box was given in terms of its x and y coordinates on the document image. Each region was also given a label from the set of 13 labels we defined. The bounding box coordinates were defined by page segmentation using the Tesseract OCR engine[2] while the labeling of the regions was done manually. Tesseract OCR performs an automatic full page segmentation of the document image thereby producing the bounded regions in the document. We allowed for manual correction of the regions by the annotators in case of a faulty or overlapping region. In total, 5 non-domain experts took part in annotating the sample of 400 document images independently. Each document image was annotated by 3 annotators (fixed number).

To make the manual annotation effort easier for the annotators, we split the 400 documents into 40 groups i.e. 10 documents per group, so that they had the liberty to annotate a minimum of 10 documents and a maximum of 400 documents. We set up the process by providing the annotators with a simple image editor tool to manually correct the segmentation (by specifying imprecise region boundaries using a variety of drawing modes such as using rectangles or arbitrary polygons) and label each region in a document image. We pre-loaded the labels into a drop-down editor to improve annotation efficiency. Hence, the annotator only needed to select the labels from a drop-down. To ensure that the annotators understood the annotation task, we provided a user guide containing complete instructions on how to use the image editor tool and carry out the labeling of the regions.

We measured the Inter-Annotator Reliability

[2] github.com/tesseract-ocr/tesseract accessed 2019-06-09

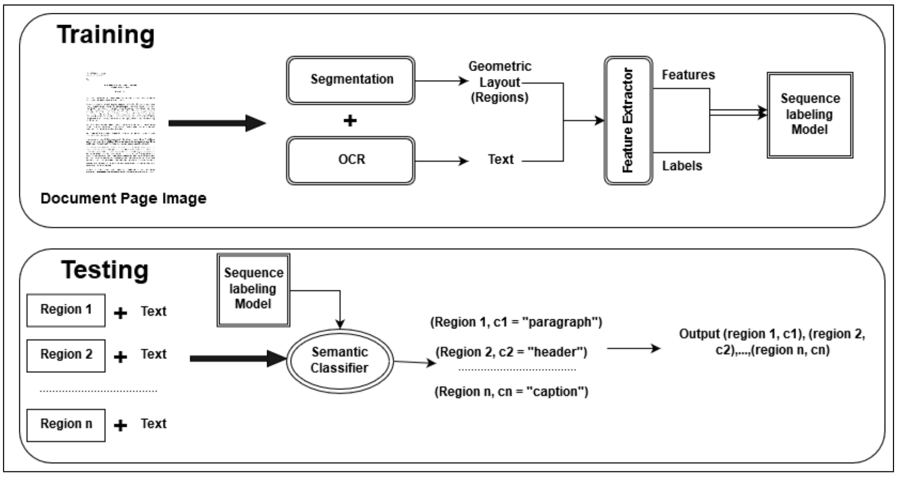


Figure 2: Implementation architecture, showing training and testing phases including the input and output for the sequence learning models

(IAR) of agreement using the Fleiss' Kappa measure (Fleiss, 1971). It has been shown to be more suitable to measure IAR when more than 2 annotators are involved, compared to other measures such as Cohen Kappa.[3] The Fleiss' Kappa value measured for our annotation task was 0.52. This value indicates *moderate agreement* between the annotators, going by the table given in (Landis and Koch, 1977) for interpreting Fleiss' Kappa values. After annotation, the main author of this paper reviewed the annotations and resolved the disagreements between the three annotators for each document image. Disagreements were resolved by majority voting and in instances where each annotator had unique annotations, the author revisited the annotated samples and made the most logical choice of label to form the gold-standard set.

4 System Architecture

Figure 2 summarizes the architecture of our semantic labeling system. During the training process, we run all input document images through the Tesseract OCR software to obtain raw text data as well as geometric layout information. The feature extractor utilizes both the layout information and raw text, when available, to produce features which go through the sequence labeling trainer together with corresponding manually labeled data, to produce the learned models. The trainer learns to assign a semantic label to the segmented regions R of a document image D. Most of the document images contain single-column layouts, hence we order the segmented regions as a sequence, from the top of the document page to the bottom. Each region $R_i \in R$ is bounded by a bounding box $B_i \in B$ that includes coherent text content and each bounding box is a set of pixels between its top left corner and bottom right corner coordinates. None of the bounding boxes overlap the other.

During testing, we want to assign a label $L_i \in$ W : $i = \{1,...,n\}$ to each region R_i. Given a sequence of regions $x = (x_1, x_2,..., x_n)$ in a document image, the task is to determine a corresponding sequence of labels $y = (y_1, y_2,..., y_n)$ for x. This can be seen as an instance of a sequence labeling problem, which attempts to assign labels to a sequence of observations. We take into account the contextual information for each of the regions in the sequence i.e. the labels of preceding or following regions are taken into account for label classification.

5 Methods

In this section, we present the sequence labeling methods for semantic labeling of document images and the evaluation procedure.

5.1 Linear-Chain CRF (LC-CRF)

CRFs are probabilistic models used to segment and label sequential data. They are reported to be very effective for semantic structure detection (Peng and McCallum, 2004; Luong et al., 2010). An inherent merit of the CRF model to perform this task is its ability to combine two classifiers: a local classifier which assigns a label to the region based on local features and a contextual classifier to model contextual correlations between adjacent regions.

[3] Fleiss' Kappa works for any number of annotators giving categorical ratings, to a fixed number of items

Feature set	Description
Without OCR	
Block coordinates	The location of the region bounding box within the document image (x and y coordinates)
Height	Normalized height of block
Width	Normalized width of block
Area	Normalized area of block
Aspect ratio	Width/height of block
Vertical position	Vertical position of region in the image (top, middle, bottom)
With OCR	
Digit	Binary feature indicating if the text in the region consists of digits or contains digits
Capital letters	Binary feature indicating if if the text in the region is all in capital case or contains capital letters
Nr of tokens	The number of tokens in a region block
Nr of lines	Binned number of lines in a region block (small, medium, large bins)
List item pattern	Binary feature indicating if text contains bullet items
Caption pattern	contains caption keywords (table, source, fig., figure)
Email keywords	Keywords found in different parts of an email
Has multi-white space (table feature)	Binary feature indicating if bounded region contains multiple white spaces between tokens.
% of white space (table feature)	The sum of white space lengths divided by the line length
Avg white space length (table feature)	The mean length of the white spaces within a line.

Table 2: Features used by the CRF methods.

Linear-chain CRFs are one well known type of CRFs which are similar to Hidden Markov Models but are reported to perform better (Peng and McCallum, 2004). They have one chain of connected labels. As CRF is a feature-based method, we implement two models with different feature sets in our work (see Table 2). We use the scikit-learn Python package, sklearn-crfsuite for implementation of our CRF models.

LC-CRF without OCR (LC-CRF$_1$): In this model, we exclude any features that can be extracted from the OCR output. That is, we consider only geometric/physical layout features to predict the label of a region in a document. The LC-CRF classifier will learn regions based on their position and location on the bounding box level of the document image. For example, it is common for *titles* to appear at the top of documents so the model may learn this observation from the extracted features.

LC-CRF with OCR (LC-CRF$_2$): By virtue of the generality and flexibility of CRF model, it is promising to achieve better performance by extending feature sets and exploring higher-level dependencies (Shetty et al., 2007). Luon et al. (2010) and Yang et al. (2017) report that by adding textual information to their models, there was an improvement in performance. We implement another LC-CRF model extending the feature set by including textual features from the OCR output. We also consider features for detecting tables. We re-use a subset of features for table detection in (Ghanmi and Abdel, 2014).

5.2 Recurrent Neural Networks (RNNs)

RNNs are a class of nets that are used for sequence learning. They can simultaneously take a sequence of inputs and produce a sequence of outputs. We transform the extracted feature sets of the CRF models into a 3D tensor and use this as input to the network. The shape of the 3D tensor is the number of input samples, the number of sequence regions per input sample and the number of features per sequence region. Therefore a shape of (300, 20, 30) indicates an input tensor of 300 document page samples, 20 regions per sample and 30 features for each region.

We use a Bidirectional-LSTM architecture for our network. Two neural models (RNN$_1$ and RNN$_2$) are trained and evaluated as such implemented for the CRF models, using feature sets with and without OCR features. Hyper-parameters are set in reference to the best performing configurations in Reimers and Gurevych (2017) with minor deviations. We use the adam algorithm for gradient descent optimization (Kingma and Ba, 2015). We don't include an embedding layer since we deal with numerical inputs, and set the number of recurrent units to 100 for all 3 hidden layers. Kernel and recurrent (l2) regularizers are added to our input layer. We introduce a batch normalization layer before the input layer and before each hidden layer to normalize the input values for our network. Normalizing or scaling the input values to a standard scale helps the network to learn the optimal parameters for each input node quickly and therefore, quickly find the minimum loss. Batch normalization also helps to improve the convergence properties of the network, has the effect of accelerating the training process of the network, and in some cases improves the performance of

	LC-CRF$_1$	LC-CRF$_2$	RNN$_1$	RNN$_2$
Overall Micro F_1	0.736	0.851	0.775	**0.855**
table	0.667	**0.897**	0.708	0.877
paragraph	0.617	**0.811**	0.622	0.774
page number	0.946	**0.966**	0.913	0.936
list item	0.336	0.594	0.559	**0.697**
heading	0.564	**0.706**	0.584	0.619
page header	0.868	**0.914**	0.846	0.865
title	0.571	0.703	0.677	**0.747**
footer	0.781	0.860	0.855	**0.868**
caption	0.667	0.742	0.742	**0.771**
email header	0.907	0.972	0.944	**0.991**
email body text	0.944	0.972	0.962	**0.989**
email signature	0.935	**0.987**	0.969	0.982
email footer	0.969	0.974	0.979	**1.000**

Table 3: Comparative performances among LC-CRF$_1$, LC-CRF$_2$, RNN$_1$ and RNN$_2$ models for semantic labeling. Category-specific performance given in F1. Results in bold mark the best system for each category.

the model. The inclusion of batch normalization layers in our network proves to be critical as it significantly improves performance. We add dropout regularization with a value of 0.1 to each hidden layer and use a batch size of 32 to control how often the weights of the network are updated. Furthermore, if the training loss does not decrease for 3 epochs, the learning rate is reduced by a 0.8 factor. Training is stopped if the minimum change in validation loss is less than 10^{-5} for 8 epochs or when 100 epochs are reached. We use the keras deep learning library running on top of tensorflow, for implementation of our RNN models.

5.3 Evaluation

The aim of our evaluation is to compare how sequence labeling methods perform for the task of semantic labeling of document regions and compare how their performances change with an extended feature set. We also evaluate the generalizability of our methods to a different document domain. Overall results are evaluated using the micro-averaged F_1 measure, the average of the results of 3 runs is reported per experiment. We split our dataset into train/test sets with a 70/30 ratio. We also perform 3-fold cross validation on the train set to tune the hyper-parameters of the model.

6 Results

6.1 Semantic Labeling of SemLab Dataset

Table 3 shows an overview of the results of our models comparison on the training dataset. The LC-CRF model without OCR output (LC-CRF$_1$) performs fairly well, approaching an F_1 score of 0.74. It is clear however that including features from the OCR output has a significant impact: the LC-CRF$_2$ model with OCR increases micro-averaged F_1 to 0.85. We observe that including features from the OCR output also improves performance for the RNN method, with the RNN$_2$ model gaining a 0.8 increase compared to the RNN$_1$ micro-averaged F_1 score of 0.78. When contrasting the implemented methods, we see that the RNN method performs better than the LC-CRF method on both model variations. RNN$_1$ shows better F_1 scores than the LC-CRF$_1$ on the majority of the categories and the overall micro F_1. The RNN$_2$ model also outperforms the LC-CRF$_2$ on most of the categories including the overall score. In addition, we make the following observations.

We observe that *list items*, *titles* and *headings* have the lowest scores for the best performing model. These categories usually have very similar features. For example, headings and list items are often started with numbering. Titles and headings also usually contain similar features such as having all capital letters. We also observe that list items have lower F_1 scores without OCR features. The classifier is able to only learn geometric and positional features of this category and misclassifies a lot of its samples as paragraph since both have similar locations on a document image and more so, paragraph is the majority category. The email related categories generally have high F_1 scores irrespective of the local feature sets included. This is because of the ability of sequence labeling methods to take into account the neighborhood of items; for example, an email body text is very likely to appear after an email header and thus the classifier learns this contextual knowledge.

6.2 Comparison across different document domain

In many real life scenarios, the datasets available to train models for the semantic labeling task are mainly homogeneous document images with similar or comparable layout and format. This raises the question about how generalizable a model that has been trained on a set or related set of document images is, to different domains. We trained the sequence labeling methods on our SemLab dataset which contains documents from multiple domains and tested each model on the records from the

	Testing Domain	
Method	SemLab	PRIMA
LC-CRF$_1$	0.861	0.696
LC-CRF$_2$	0.923	0.743
RNN$_1$	0.888	0.701
RNN$_2$	0.890	0.747

Table 4: Review of the transfer learning experiment. Each method is trained on the SemLab dataset and tested on in-domain and cross-domain documents. All scores are micro-averaged F_1 scores.

PRIMA dataset which contains documents from the magazine domain, not represented in our own dataset. For fair comparison, we evaluated only labels applicable to both datasets i.e. intersecting labels (header, paragraph, section heading, caption, page number, footer). For this reason we excluded some features from the 'With OCR' feature set that are directly related to the excluded labels.

Table 4 provides a summary of the performance of each method on the different domains. The results show that the methods have lower performances when evaluated on unseen data of a different domain than the training data. Both LC-CRF and RNN methods perform better when OCR information is included for the cross domain experiment. This proves that the inclusion of textual features also aids generalizability of methods across new domains for semantic labeling. Furthermore, we observe that both RNN methods are able to generalize better than the LC-CRF methods, though with slight improvements. This could be explained by the techniques specifically employed to reduce overfitting and improve generalizability power in the RNN such as the use of dropout, early stopping, l2 regularization, among others.

7 Conclusion and Future Work

In this work we have presented a comparison between state-of-the-art sequential learning models applied to the task of semantic labeling of document regions. We constructed a novel evaluation dataset to benchmark model performance on. The experimental results reveal that both methods are able to perform the task well using only a small amount of training data; with the RNN method slightly outperforming the LC-CRF method. Also, including OCR information in the feature set is promising to achieve better performance as it reduces confusion between ambiguous semantic classes. In addition, its inclusion might positively affect generalization performance, as shown by our transfer learning experiments on the PRIMA domain.

Future work includes extending the document dataset in terms of size and variety to cover more document spaces, domains and classes. Models can exploit these characteristics to better generalize to new domains. By virtue of neural networks' great power to learn latent features, we believe more (varying) data will also contribute to improving the performance levels of our neural method. An extension of the feature set used in this work could also be beneficial in improving performance scores for the implemented models.

References

[Adnan and Ricky2011] Amin Adnan and Shiu Ricky. 2011. Page segmentation and classification utilizing bottom-up approach. *International Journal of Image and Graphics*, 01.

[Antonacopoulos et al.2006] A. Antonacopoulos, D. Karatzas, and D. Bridson. 2006. Ground truth for layout analysis performance evaluation. In *Document Analysis Systems VII*, pages 302–311, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Antonacopoulos et al.2009] A. Antonacopoulos, D. Bridson, C. Papadopoulos, and S. Pletschacher. 2009. A realistic dataset for performance evaluation of document layout analysis. In *2009 10th International Conference on Document Analysis and Recognition*, pages 296–300, July.

[Antonacopoulos1998] A Antonacopoulos. 1998. Page segmentation using the description of the background. *Computer Vision and Image Understanding*, 70(3):350–369.

[Chen and Blostein2007] Nawei Chen and Dorothea Blostein. 2007. A survey of document image classification: problem statement, classifier architecture and performance evaluation. *International Journal of Document Analysis and Recognition (IJDAR)*, 10(1):1–16.

[Clausner et al.2011] C. Clausner, S. Pletschacher, and A. Antonacopoulos. 2011. Scenario driven in-depth performance evaluation of document layout analysis methods. In *2011 International Conference on Document Analysis and Recognition*, pages 1404–1408, Sep.

[Fleiss1971] Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

[Gatos et al.1999] B. Gatos, S. L. Mantzaris, K. V. Chandrinos, A. Tsigris, and S. J. Perantonis. 1999.

Integrated algorithms for newspaper page decomposition and article tracking. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR '99 (Cat. No.PR00318)*, pages 559–562, Sep.

[Ghanmi and Abdel2014] Nabil Ghanmi and Belaïd Abdel. 2014. Table detection in handwritten chemistry documents using conditional random fields. In *ICFHR*, pages p. 146–151, Crete, Greece.

[Kamola et al.2015] Grzegorz Kamola, Michal Spytkowski, Mariusz Paradowski, and Urszula Markowska-Kaczmar. 2015. Image-based logical document structure recognition. *Pattern Anal. Appl.*, 18(3):651–665.

[Kim et al.2000] Jongwoo Kim, Daniel X. Le, and George R. Thoma. 2000. Automated labeling in document images. In *Document Recognition and Retrieval*.

[Kingma and Ba2015] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

[Kise et al.1998] Koichi Kise, Akinori Sato, and Motoi Iwata. 1998. Segmentation of page images using the area voronoi diagram. *Comput. Vis. Image Underst.*, 70(3):370–382.

[Landis and Koch1977] J Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74.

[Luong et al.2010] Minh-Thang Luong, Min-Yen Kan, and Thuy Dung Nguyen. 2010. Logical structure recovery in scholarly articles with rich document features. *Int. J. Digit. Library Syst.*, 1(4):1–23.

[Mao et al.2003] Song Mao, Azriel Rosenfeld, and Tapas Kanungo. 2003. Document structure analysis algorithms: a literature survey. In *Document Recognition and Retrieval X, Santa Clara, California, USA, January 22-23, 2003, Proceedings*, pages 197–207.

[Marinai2008] Simone Marinai, 2008. *Introduction to Document Analysis and Recognition*, pages 1–20. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Pavlidis and Zhou1992] Theo Pavlidis and Jiangying Zhou. 1992. Page segmentation and classification. *CVGIP: Graph. Models Image Process.*, 54(6):484–496.

[Peng and McCallum2004] Fuchun Peng and Andrew McCallum. 2004. Accurate information extraction from research papers using conditional random fields. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2004, Boston, Massachusetts, USA, May 2-7, 2004*, pages 329–336.

[Rangoni et al.2012] Yves Rangoni, Abdel Belaïd, and Szilárd Vajda. 2012. Labelling logical structures of document images using a dynamic perceptive neural network. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 45–55.

[Reimers and Gurevych2017] Nils Reimers and Iryna Gurevych. 2017. Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. *CoRR*, abs/1707.06799.

[Shetty et al.2007] Shravya Shetty, Harish Srinivasan, Sargur Srihari, and Matthew Beal. 2007. Segmentation and labeling of documents using conditional random fields. *Proceedings of SPIE - The International Society for Optical Engineering*, 6500:6500–1.

[Siegel et al.2018] Noah Siegel, Nicholas Lourie, Russell Power, and Waleed Ammar. 2018. Extracting scientific figures with distantly supervised neural networks. *CoRR*, abs/1804.02445.

[Stahl et al.2018] Christopher Stahl, Steven Young, Drahomira Herrmannova, Robert Patton, and Jack Wells. 2018. Deeppdf: A deep learning approach to extracting text from pdfs. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

[Tao et al.2013] Xin Tao, Zhi Tang, and Canhui Xu. 2013. Document page structure learning for fixed-layout e-books using conditional random fields. *Proceedings of SPIE - The International Society for Optical Engineering*, 9021.

[Todoran et al.2005] Leon Todoran, Marcel Worring, and Arnold W. M. Smeulders. 2005. The uva color document dataset. *International Journal of Document Analysis and Recognition (IJDAR)*, 7(4):228–240.

[Yang et al.2017] X. Yang, E. Yumer, P. Asente, M. Kraley, D. Kifer, and C. L. Giles. 2017. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4342–4351.