# Information Extraction From Structured, Semi-structured and Unstructured Banking Documents

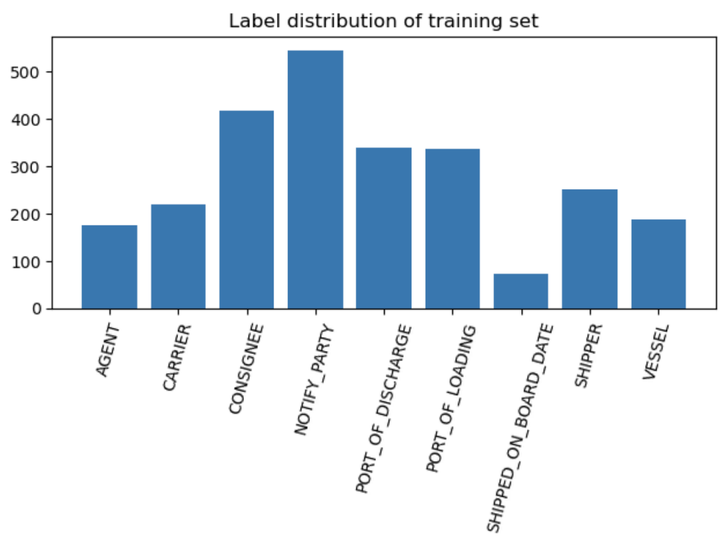
Furkan Pala – 150180109
Advisor: Assoc. Prof. Dr. Gülşen Cebiroğlu Eryiğit

## Introduction

**Information extraction** is the automatic retrieval of the interested information from different channels such as text and scanned document images. In this project, named-entity recognition (NER) which is a popular subtask of information extraction in NLP is adapted to retrieve the essential information from banking documents using multimodal deep learning methods.

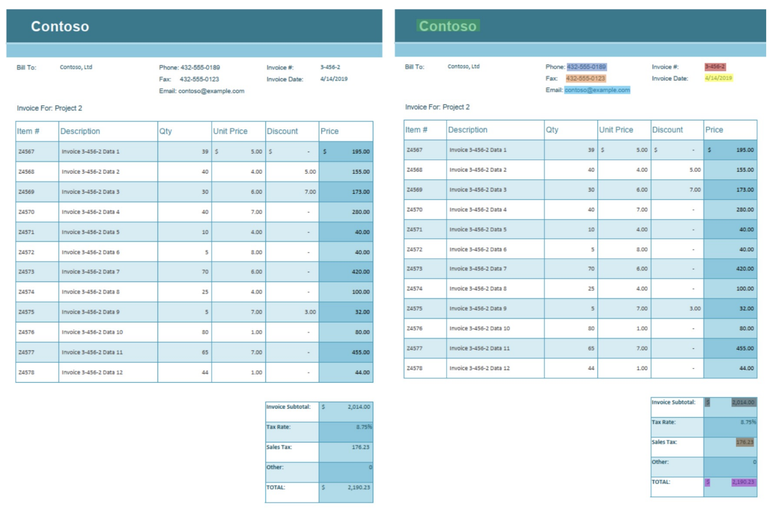
## Dataset

Our in-house dataset provided by Yapı Kredi Technology consists of bill of lading documents which are the invoices about the goods in transportation. We have 133 documents for training, 29 for validation and 29 for testing. Each word was annotated by human annotators with one of the 10 labels including the OTHER label for non-interested words. Distribution of the related labels in training set is as follows:




*Key entity label distribution in training set*

Due to the privacy issues, samples from the bill of lading dataset cannot be shared but, just for the demonstration, an example document with annotation is provided in ViBERTgrid paper as seen below




*An example document with annotation of key entity labels such as company name, address and total amount*

### Public Datasets

Because of the sensitive information in banking documents and invoices, there is scarcity of publicly available datasets in this domain. Receipts, on the other hand, carry some similarity to invoices and have a publicly available and annotated dataset, namely Scanned Receipts OCR and Information Extraction (SROIE) in ICDAR 2019 competition. An example receipt can be seen below:

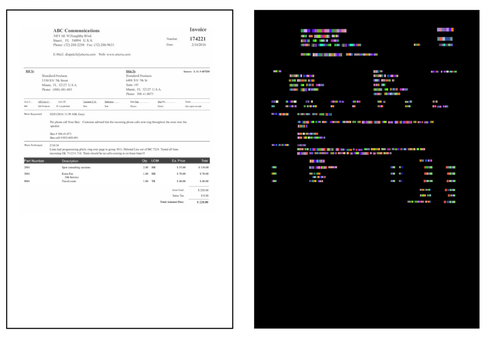

*An example receipt with annotation of key entity labels: company name, date, address and total amount. Needs further pre-processing*

## Methods

While pure text-based models do not capture the visual aspects and layout such as the position of the textual elements, pure vision-based models lack the meaning of the text in the natural language. Thus, multimodal analysis plays a key role in terms of understanding both the layout and linguistic information of the document. To extract the text from scanned document images, an OCR engines was used at the beginning. Three recent studies were investigated in this project:

### Chargrid

Chargrid represents a document as a 2D grid of characters so that the spatial information is preserved. Each pixel occupied by a character is substituted by the corresponding code for that character; thus, information extraction task is interpreted as a pixel-wise semantic segmentation over the Chargrid tensor.



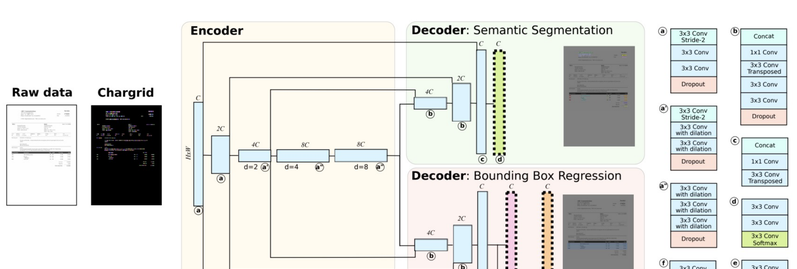
*A raw document (left) and generated Chargrid tensor (right)*

### BERTgrid

BERTgrid is an advancement of Chargrid with the idea of encoding each pixel inside a key information region in the document with a contextualized word embeddings generated using BERT language model instead of encoding each character individually.

*An example invoice (left) with ground-truth segmentation mask (middle) and generated BERTgrid tensor (right)*

Deep neural network used in BERTgrid is the same as in Chargrid composed of one encoder and two decoders of which we only use the semantic segmentation branch since the goal is to extract information in terms of classifying each word with the correct entity label, i.e., NER task.



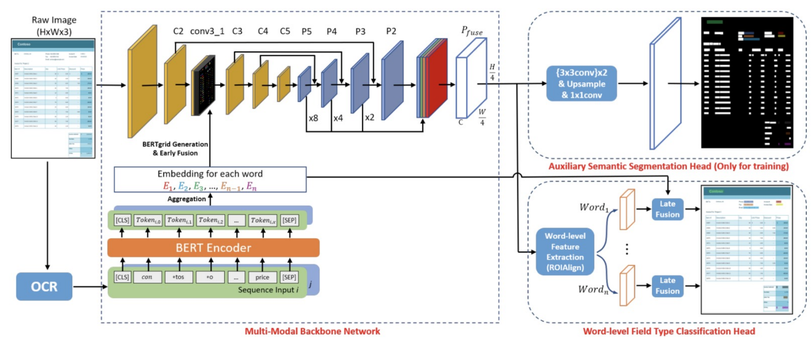

*Chargrid architecture which is also used in BERTgrid*

### ViBERTgrid

BERTgrid has two downsides:

- Document image which carries significant visual features is not fed into the deep neural network
- BERT language model is not fine-tuned during the training

ViBERTgrid addresses the above issues by using ResNet18-FPN as CNN backbone to conduct early fusion on BERTgrid tensor and fine-tuning both BERT and CNN with a joint training strategy. Architecture of ViBERTgrid can be seen below:




*ViBERTgrid architecture*

## Results

Since NER is a multi-class classification task, we evaluate the aforementioned models with respect to the F1 scores on word level. ViBERTgrid outperforms all the other models including some baselines such as BERT on serialized text and LayoutLMv2 as seen below:

| Label | BERT base | LayoutLMv2 | BERTgrid base | BERTgrid banking | ViBERTgrid |
|---|---|---|---|---|---|
| Macro | 52.31 | 62.92 | 70.34 | 68.31 | **74.42** |
| Micro | 63.35 | 75.11 | 75.84 | 76.14 | **80.13** |

*Micro and macro F1 scores of the models on the bill of lading dataset*

Detailed classification results for ViBERTgrid on bill of lading dataset can be seen below:

| Label | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| AGENT | 29.73 | 40.74 | 34.38 | 27/176 |
| CARRIER | 48.48 | 71.11 | 57.66 | 45/220 |
| CONSIGNEE | 93.83 | 100 | 96.82 | 76/417 |
| NOTIFY PARTY | 90.4 | 91.87 | 91.13 | 123/546 |
| PORT OF DISCHARGE | 78.26 | 81.82 | 80 | 66/340 |
| PORT OF LOADING | 73.61 | 71.62 | 72.6 | 74/338 |
| SHIPPED ON BOARD DATE | 62.5 | 83.33 | 71.43 | 18/72 |
| SHIPPER | 92 | 90.2 | 91.09 | 51/251 |
| VESSEL | 70.45 | 79.49 | 74.7 | 39/188 |
| Macro | 71.03 | 78.91 | 74.42 | |
| Micro | 77.86 | 83.04 | 80.13 | |

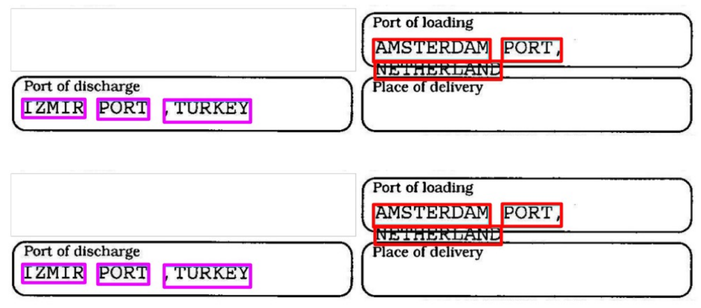
*Detailed classification results for ViBERTgrid on bill of lading dataset. Support column represents test/train supports*

ViBERTgrid works well when it is trained on documents with different layouts as well, as seen in the below table explains the results on the SROIE dataset:

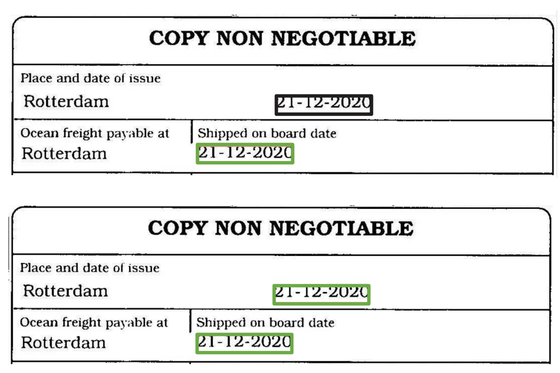
| Label | Precision | Recall | F1-Score | Test Support |
|---|---|---|---|---|
| ADDRESS | 91.00 | 91.00 | 91.00 | 79 |
| COMPANY | 94.00 | 98.00 | 96.00 | 181 |
| DATE | 88.00 | 98.00 | 93.00 | 46 |
| TOTAL | 92.00 | 95.00 | 93.00 | 111 |
| Macro | 92.17 | 95.92 | 94.01 | |

*ViBERTgrid results on the SROIE dataset*

The figures below depict some visual classification results even though they are cropped from the whole document to not reveal the sensitive information:




*Correct classification of the prediction (below) with respect to the ground-truth annotation (above). Color codes are port of discharge and port of loading.*




*Misclassification of the prediction (below) with respect to the ground-truth annotation (above). Color codes are shipped on board date and other.*

## Conclusion

This project was an extensive study to investigate the performance of the multimodal models on information extraction in banking domain which is still an active area of research. I implemented state-of-the-art papers and made contribution to our NLP team in Yapi Kredi Technology. There are still gaps in the literature due to scarcity of public datasets; thus, further studies including providing multimodal datasets are surely welcomed in the area. Also, since computer vision techniques are developed quite rapidly in parallel with NLP, there is always a possibility to transform and combining the newly emerged methods in information extraction.

## References

1. A. R. Katti, C. Reisswig, C. Guder, S. Brarda, S. Bickel, J. H¨ohne, and J. B. Faddoul, “Chargrid: Towards understanding 2d documents,” 2018.
2. T. I. Denk and C. Reisswig, “Bertgrid: Contextualized embedding for 2d document representation and understanding,” 2019.
3. W. Lin, Q. Gao, L. Sun, Z. Zhong, K. Hu, Q. Ren, and Q. Huo, “Vibertgrid: A jointly trained multi-modal 2d document representation for key information extraction from documents,” 2021. [Online]. Available: https://arxiv.org/abs/2105.11672.

İTÜ
ISTANBUL TECHNICAL UNIVERSITY
Department of Computer Engineering
Faculty of Computer and Informatics Engineering
İTÜ Ayazaga Campus, 34469, Maslak/Istanbul
bbf@itu.edu.tr
http://bm.itu.edu.tr