# A Deep Learning-Based Approach for Named Entity Recognition on Commercial Receipts

THE UNIVERSITY OF AUCKLAND NEW ZEALAND
SCHOOL OF COMPUTER SCIENCE

Fraedom®
VISA

Author: Spencer Han | Academic Supervisor: Sebastian Link | Industry Advisor: David Duan (Principle Data Scientist), Casper Hart (Data Scientist) @ Fraedom Ltd. | Date: 27/11/2020

## 1 Introduction

Transaction management and expense analytics service providers increasingly seek to establish an automated system to harness information from commercial receipts.

With the recent enhancement in the area of optical character recognition, the transformation from digital invoice documents to invoice text data is made possible and reliable. It is now becoming a priority to extract and identify essential entity information from the transformed invoice text data.

In this affiliated study with Fraedom Limited, we proposed **a deep learning-based named entity recognition method** to identify named entities from Amazon Textract processed commercial receipts text data.

### Research Questions:

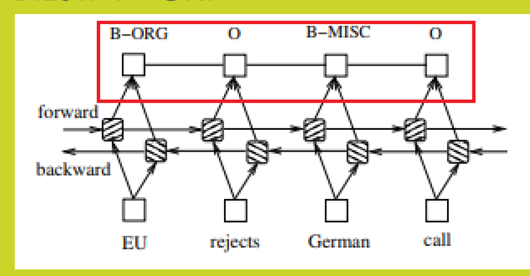
1. Are deep-learning models capable of extracting entity information from unstructured text data?
2. Which is the best performing model architecture using our data?
3. What's the impact of optical character recognition(OCR) noises?
4. What's the future improvement in this area?

## 2 Methodology

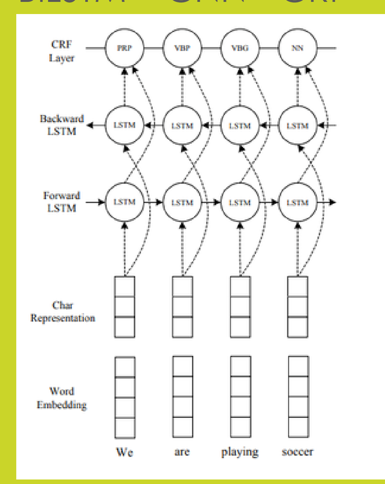
We begin with preprocessing the transformed invoice text data to create labelled training, validation and testing dataset. And then we implemented **three** deep learning architectures [1][2][3] based on the concept of Bi-directional Long Short-term Memory Networks, Convolutional Neural Networks, Conditional Random Fields [4] and Softmax [5] using Python. We then evaluated the three architectures according to their training performance against our validation dataset and testing dataset. Lastly, we provide a working example of solving a named entity recognition task using our best model as well as recommendations for future work in this area.

### Three Models:

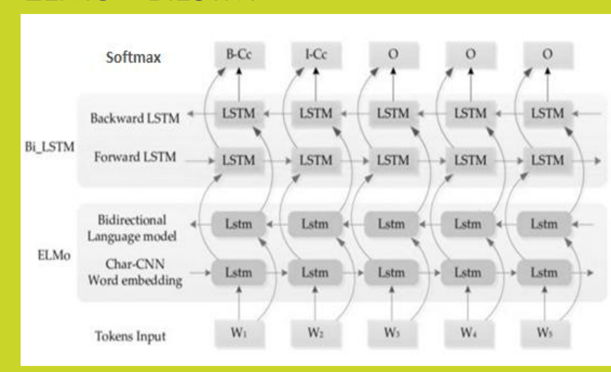
BiLSTM - CRF



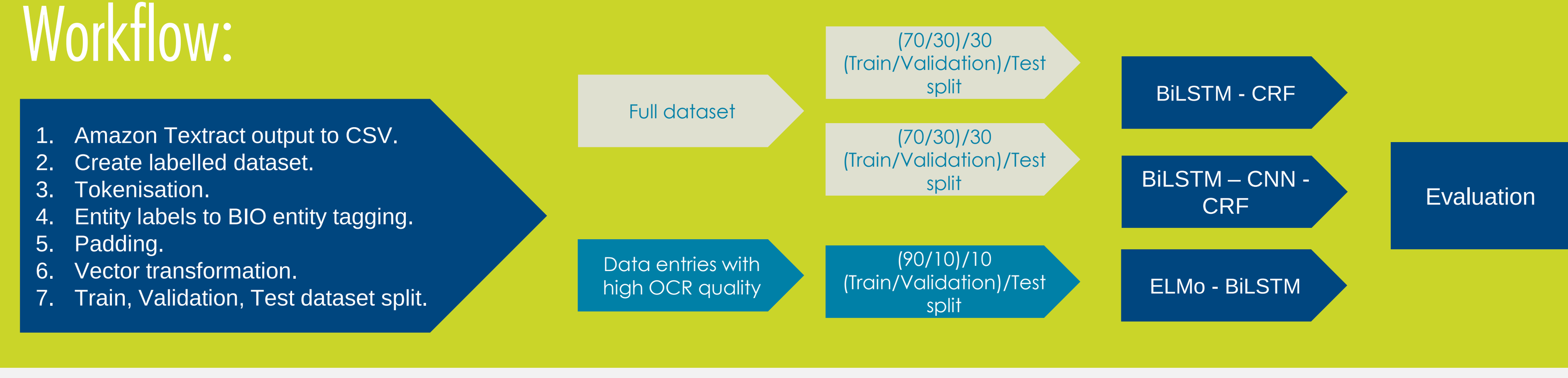

BiLSTM – CNN - CRF




ELMo - BiLSTM




### Workflow:




## 3 Results

1. Trained with all available data (split 9:1 between (train + validation) and test set, then 9:1 between train and validation set)

BiLSTM - CRF

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| AMOUNT | 0.88 | 0.88 | 0.88 | 391 |
| DATE | 0.61 | 0.30 | 0.40 | 183 |
| GST | 0.47 | 0.51 | 0.49 | 35 |
| MERCHANT | 0.52 | 0.41 | 0.46 | 194 |
| TIME | 0.80 | 0.37 | 0.50 | 123 |

BiLSTM – CNN - CRF

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| AMOUNT | 0.93 | 0.90 | 0.92 | 476 |
| DATE | 0.61 | 0.38 | 0.47 | 215 |
| GST | 0.38 | 0.55 | 0.45 | 38 |
| MERCHANT | 0.55 | 0.45 | 0.50 | 223 |
| TIME | 0.93 | 0.40 | 0.56 | 135 |

ELMo - BiLSTM

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| AMOUNT | 0.94 | 0.95 | 0.95 | 390 |
| DATE | 0.84 | 0.88 | 0.86 | 180 |
| GST | 0.80 | 0.99 | 0.89 | 35 |
| MERCHANT | 0.60 | 0.54 | 0.57 | 193 |
| TIME | 0.96 | 0.98 | 0.97 | 121 |

2. Trained with all available data (split 7:3 between (train + validation) and test set, then 7:3 between train and validation set)

BiLSTM - CRF

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| AMOUNT | 0.92 | 0.88 | 0.90 | 1330 |
| DATE | 0.47 | 0.27 | 0.34 | 562 |
| GST | 0.50 | 0.56 | 0.53 | 117 |
| MERCHANT | 0.53 | 0.41 | 0.47 | 613 |
| TIME | 0.59 | 0.23 | 0.33 | 360 |

BiLSTM – CNN - CRF

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| AMOUNT | 0.91 | 0.87 | 0.89 | 1330 |
| DATE | 0.56 | 0.26 | 0.36 | 562 |
| GST | 0.43 | 0.50 | 0.46 | 117 |
| MERCHANT | 0.50 | 0.38 | 0.43 | 613 |
| TIME | 0.79 | 0.33 | 0.47 | 360 |

ELMo - BiLSTM

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| AMOUNT | 0.95 | 0.92 | 0.93 | 1329 |
| DATE | 0.85 | 0.76 | 0.80 | 561 |
| GST | 0.84 | 0.95 | 0.89 | 117 |
| MERCHANT | 0.57 | 0.58 | 0.58 | 612 |
| TIME | 0.97 | 0.97 | 0.97 | 359 |

3. Trained with data with high OCR quality per invoice (split 9:1 between (train + validation) and test set, then 9:1 between train and validation set)

BiLSTM - CRF

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| AMOUNT | 0.89 | 0.84 | 0.86 | 377 |
| DATE | 0.68 | 0.32 | 0.44 | 186 |
| GST | 0.38 | 0.48 | 0.42 | 31 |
| MERCHANT | 0.55 | 0.50 | 0.52 | 150 |
| TIME | 0.91 | 0.44 | 0.59 | 112 |

BiLSTM – CNN - CRF

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| AMOUNT | 0.94 | 0.84 | 0.89 | 377 |
| DATE | 0.78 | 0.32 | 0.45 | 186 |
| GST | 0.68 | 0.61 | 0.64 | 31 |
| MERCHANT | 0.57 | 0.48 | 0.52 | 150 |
| TIME | 0.86 | 0.45 | 0.59 | 112 |

ELMo - BiLSTM

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| AMOUNT | 0.95 | 0.93 | 0.94 | 377 |
| DATE | 0.87 | 0.91 | 0.89 | 184 |
| GST | 0.81 | 0.97 | 0.88 | 31 |
| MERCHANT | 0.61 | 0.48 | 0.54 | 149 |
| TIME | 0.96 | 0.99 | 0.97 | 110 |

### Live Prediction example with ELMo - BiLSTM:

```
inputInvoice = 'OUTDOOR, CONCEPTS, Tax Invoice, Outdoor Concepts Ltd, GST
Number 069-712-746, TIME, 11:12PM, Date,  14 Oct 2019, New Zealand,(NZD) Sub
Total, 6539, WEBER PULSE, 1, $349.00, $349.00, 1, $799.00, $799.00, 2000,
7181, Weber Pulse, 1, $69.95, $69.95, $0.00, Premium Cover, 1000/2000, 6415,
Weber Small Drip, 1, $14.95, $14.95, $0.00, Pan, Product Cost:, $1,148.00,
Delivery Details:, Local Oversize $50.00, Sub Total:, $1,198.00, GST:,
$156.26, Tax Invoice Total:, (NZD) $1,198.00, Payments, Method, Ref, Amount,
Total Paid:, (NZD) $1,198.00, 14 Oct 2019, ShopifyV2 - shopify_payments,
$1,198.00, Outstanding:, (NZD) $0.00, Terms: Payment is due before delivery
of goods. For customers on account, invoices are to be paid no
20th of the month following receipt of invoice.'
```

```
outdoor , concepts , tax invoice , outdoor concepts ltd , gst number 069 - 712 - 7
46 , time , 11:12pm , date ,   14 oct 2019 , new zealand,(nzd ) sub total , 6539 ,
weber pulse , 1 , $ 349.00 , $ 349.00 , 1 , $ 799.00 , $ 799.00 , 2000 , 7181 , we
ber pulse , 1 , $ 69.95 , $ 69.95 , $ 0.00 , premium cover , 1000/2000 , 6415 , we
ber small drip , 1 , $ 14.95 , $ 14.95 , $ 0.00 , pan , product cost : , $ 1,148.0
0 , delivery details : , local oversize $ 50.00 , sub total : , $ 1,198.00 , gst :
, $ 156.26 , tax invoice total : , ( nzd ) $ 1,198.00 , payments , method , ref ,
amount , total paid : , ( nzd ) $ 1,198.00 , 14 oct 2019 , shopifyv2 - shopify_pay
ments , $ 1,198.00 , outstanding : , ( nzd ) $ 0.00 , terms : payment is due befor
e delivery of goods . for customers on account , invoices are to be paid no later
than the 20th of the month following receipt of invoice .
```

TensorFlow
Keras
spaCy

## 4 DISCUSSION

1. The NER result for 'Merchant' entity is the worst.
2. The ELMo – BiLSTM is the best performing model.
3. The ELMo – BiLSTM consumes the most resource and time to train.
4. The Labelling process takes the most amount of effort to complete compare to other tasks in this project. Semi-supervised and unsupervised learning can mitigate this issue.
5. This study does not incorporate rule-based NER system which can be beneficial in some circumstances.
6. Hyperparameter tuning was not investigated thoroughly, thus the model evaluation result may not reflect the true capabilities of these models.
7. OCR noises and quality were analysed at the sequence level rather than at the token level. This may not reflect the true impact of OCR noises per entity token.
8. The tokenisation was done using a tokeniser built for English document. It assumes the underlying natural language data follows English grammar rules which in this case, it does not.

## 5 CONCLUSION

In this study, we adopt a supervised machine learning workflow. We used a variety of Python libraries to annotate OCR'ed receipt text with our chosen named entity labels and to develop three deep learning models to achieve our NER goal: BiLSTM – CRF, BiLSTM – CNN – CRF and ELMo - BiLSTM.

The models are evaluated under three different train/validation/test split schemes. We also investigated the impact of OCR noises on the result by training the model using data contain misreadings versus training the model using data with high OCR qualities.

We discovered that BiLSTM – CRF is the least expensive one to implement but produces the worst prediction result amongst the three models. The ELMo – BiLSTM with contextualised word embedding consumes the most resources to compute but yields the best prediction result amongst the three models. Meanwhile, BiLSTM – CNN – CRF lies in the middle of the former two models. With the ELMo – BiLSTM model, we retained an above 80% precision, recall and f1-score for most named entities types except for the 'merchant' entity. We also presented a live example to demonstrate how the ELMo – BiLSTM model identifies named entities on an unlabelled receipt text. As a result, we concluded that even with the presence of OCR noises, our best performing model (ELMo – BiLSTM) is able to produce a reliable NER result and can be reused in a production environment

## Acknowledgement

I would like to take this opportunity to thank my supervisor Sebastian Link for his guidance, encouragement and support throughout my dissertation. I am also grateful to Fraedom NZ Ltd. for providing this research opportunity. A great appreciation goes to David Duan (Principle Data Scientist) and Casper Hart (Data Scientist) at Fraedom, for providing project supervision as well as insights and feedback on my research methodologies from time to time. Without their supports, I may not be able to set the correct direction for this project and overcome technical challenges that were faced during implementing natural language processing approaches.

**References:**

[1] Z. Huang, W. Xu and K. Yu, Bidirectional LSTM-CRF Models for Sequence Tagging, 2015.
[2] X. Ma and E. Hovy, End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF, 2016.
[3] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer, Deep contextualised word representations, 2018.
[4] J. D. Lafferty, A. McCallum and F. C. N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in Proceedings of the Eighteenth International Conference on Machine Learning, San Francisco, CA, USA, 2001.
[5] M. Zhang, G. Geng and J. Chen, "Semi-Supervised Bidirectional Long Short-Term Memory and Conditional Random Fields Model for Named-Entity Recognition Using Embeddings from Language Models Representations," Entropy, vol. 22, p. 252, 2 2020.

THE UNIVERSITY OF AUCKLAND
Te Whare Wananga o Tamaki Makaurau
NEW ZEALAND