International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
Impact Factor (2020): 6,733

Enhancing Data Extraction from Scanned Official Correspondences Using Named Entity Recognition: A Case Study at Kaduna Polytechnic

Akande H. F.*, Ibrahim, A. I., Muhsin A.**, Lawal R. L.****

*Department of Chemical Engineering, Kaduna Polytechnic, Kaduna, Kaduna State, Nigeria
**Department of Computer Science, Kaduna Polytechnic, Kaduna, Kaduna State, Nigeria

Abstract: ***In this research, we explore the potential of Named Entity Recognition (NER), a Natural Language Processing (NLP) component, for efficient data extraction from official correspondences at Kaduna Polytechnic, Nigeria. Leveraging Optical Character Recognition (OCR) technology, we digitised around 460 official correspondences to train a NER model using the SpaCy Python library. The dataset was split into a training set of 400 documents and a test set of 60 documents. The NER model's performance was assessed using precision, recall, and F1 score metrics. After training, the model achieved an F1 score of 0.92 on the test set, demonstrating its improved ability to predict and label named entities accurately. This study offers tangible evidence of how NLP tools like SpaCy can be utilised to enhance data management tasks in an academic environment, pointing towards broader applications in data extraction and digitisation across similar institutional settings.***

Keywords: Named Entity Recognition, Natural Language Processing, Optical Character Recognition, Deep Neural Networks

1.Introduction

The rapidly growing field of data digitisation and information retrieval has impacted various sectors, including educational institutions. In this context, like many similar institutions, Kaduna Polytechnic faces the challenge of efficiently managing and accessing data from a vast array of official correspondences such as employment, promotion, and retirement letters. Converting these documents into a digitised format allows for more efficient data management and easier access to historical records, thereby enhancing operational efficiency.

Named Entity Recognition (NER), a subfield of Natural Language Processing (NLP), has shown great potential in enhancing data extraction from unstructured text data. NER focuses on identifying and classifying named entities within text into predefined categories, such as names of persons, organisations, locations, expressions of times, quantities, and others. This study seeks to leverage the capabilities of NER for efficient data extraction from scanned documents in an academic context.
SpaCy, a high-performance Python library for NLP, offers powerful capabilities for NER tasks. While its application in various domains is well-documented, this study is pioneering in its application within the specific context of Kaduna Polytechnic's document management system.

This study, therefore, aims to apply and evaluate the effectiveness of SpaCy's NER capabilities in extracting relevant data from scanned official correspondences at Kaduna Polytechnic. Using Optical Character Recognition (OCR) technology, approximately 300 official correspondences were scanned and converted into a digital format suitable for NLP processing. The entities within these documents, including the title, sender, receiver, body text, and signatory, were annotated and labelled using the BIO (Begin, Inside, Outside) tagging format. This annotated data served as a training set for a SpaCy NER model.

This study aims to explore whether such an approach can simplify and enhance data management tasks within an academic setting. Through this endeavour, we hope to contribute to the NLP and data management field by demonstrating how robust yet straightforward tools like SpaCy can be employed effectively in real-world settings.

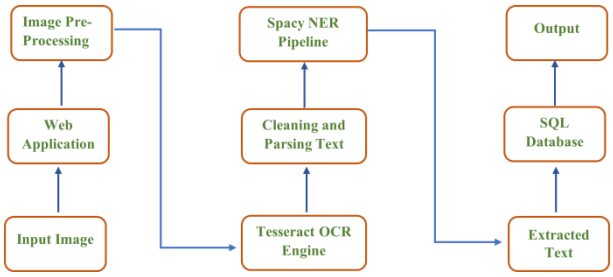

Figure 1: Proposed System Architecture

2.Review of Related Work

Named Entity Recognition (NER) has undergone significant evolution and development, transitioning from rule-based and statistical models to machine learning and deep learning approaches. Early methods, such as hybrid models integrating maximum entropy, neural networks, and pattern-selection rules (Seon et al., 2001), demonstrated potential but faced limitations in handling unknown words and scalability. Machine learning techniques, including conditional random field models (Freire et al., 2012) and Naive Bayes and SVM models (Shabat et al., 2014), showed promising results in improving NER performance. Deep learning architectures, such as deep neural networks (DNNs) (Gridach, 2016; Al-Smadi et al., 2020), recurrent neural networks (RNNs), and transformer models (Sun and Li, 2023; Al-Qurishi and Souissi, 2021), have significantly advanced NER systems.

Despite these advancements, challenges persist in NER research. Language bias remains an issue, as most studies focus on English and Chinese, leaving low-resource

languages underexplored (Nadeau and Sekine, 2007). Efforts have been made to address this, with researchers exploring multilingual and cross-lingual NER (Ruder et al., 2018; Al-Rfou et al., 2019). Semi-supervised and unsupervised learning approaches have emerged as alternatives to reduce reliance on annotated data (Lample et al., 2016; Artetxe et al., 2018). Data augmentation techniques, such as synthetic training data generation (Kurata et al., 2016) and transfer learning with large pre-trained models like GPT-3 and BERT (Brown et al., 2020; Devlin et al., 2018), have also contributed to improved NER accuracy.

As NER progresses, new frontiers are being explored. Incorporating external knowledge bases and resources has shown promise in enhancing contextual understanding of entities (Zeng et al., 2014; Yang & Chang, 2018). Domain adaptation techniques address the challenge of applying NER models to new domains or genres (Kim et al., 2017). Real-time processing requirements have spurred the development of lightweight and efficient models (Zhou & Xu, 2015; Lan et al., 2019). However, challenges remain, such as handling entity variability in unstructured, noisy, and multilingual data and addressing the interpretability issue in deep learning models (Ribeiro et al., 2016).

In conclusion, while significant progress has been made in NER research, there is still ample opportunity for innovation. The future of NER will likely involve novel machine learning techniques, integration of external knowledge, advancements in unsupervised and semi-supervised learning, exploration of multilingual and domain-specific NER, and efforts to address challenges such as real-time processing and model interpretability.

3.Methodology

This study's methodology involves collecting a diverse dataset of 460 documents, partitioned into a training set of 400 documents and a testing set of 60 documents. Text content was extracted from these documents using the Tesseract OCR system, then annotated exhaustively using BIO tags to prepare for the Named Entity Recognition (NER) model training.
The NER model was trained using transfer learning techniques, with pre-trained models fine-tuned to optimise the recognition and classification of named entities.

The application was built using Python 3.9 and several libraries, including Tesseract OCR, OpenCV, Pandas, Spacy, Regex, and Django.

The Django-based application enables users to upload scanned document images, triggering the OCR and NER pipeline. This pipeline includes image pre-processing, noise reduction, bounding box prediction, and error handling mechanisms, converting the images into machine-readable text for the NER model.

Error handling and system troubleshooting considerations were incorporated into the application design to deal with potential issues like file format incompatibility, image quality, and server-side errors. Performance evaluation was done using metrics such as precision, recall, and F1 score on the testing set of 60 documents.

Ethical guidelines were strictly followed, with personally identifiable information either excluded or anonymised, and all data stored and processed in a secure environment. This comprehensive methodology ensures transparency and can be replicated in similar studies, thereby enhancing the scientific value and integrity of the research.

A. Modules of the Proposed System

In this study, an entity recognition system that can be accessed through a user-friendly Django Web UI was implemented. Users can easily upload scanned images of the desired document for archiving.

B. System Requirements

The system requirement contains and describes what the clients want for a particular system. It is a structural document with detailed descriptions of the system services. Some of the system requirements are:

- User should be able to upload scanned images for archiving.
- User would be able to view the result of the extraction.
- User should be able to retrieve extracted document.

C. User Requirements

Some of the user requirements for this system are:

- The system requires a scanner ready computer.
- The system will have a database to store all official documents that are currently available.

D. Functional Requirements

Some of the functional requirements for this system are:

- It should extract text from documents accurately.
- The system should be able to store the extracted text into a database.
- The system should be able to return stored text from the database.

E. Non-Functional Prerequisites

i. The system will be able to function and remain active for at least 90% of the time.
ii. The layout of the system will be focused and clear. This would help to reduce the likelihood of users being confused with the interface. It will only display information that is relevant to the current work.
iii. The system will have to deal with massive amounts of data storage and a large number of users accessing the data simultaneously.

F. Hardware and Software Prerequisites

This section stresses the hardware and software components required for running this application.

G. Hardware Requirements

- Intel Core i5 processor
- 8 GB RAM
- A scanner device

H. Software Requirements

- Python web framework (Django)
- Tesseract OCR
- Pillow (Python Imaging Library)
- OpenCV
- Tesseract OCR
- Pandas
- Spacy
- Regex
- HTML, CSS, JavaScript
- MySQL Database

4.Result and Discussion

Table 1: The Trained Entity Recognition Model metrics

E	#	LOSS TOK2VEC	LOSS NER	ENTS_F	ENTS_P	ENTS_R	SCORE
0	0	0.00	178.82	0.00	0.00	0.00	0.00
2	200	4099.11	13924.53	85.23	89.53	81.32	0.85
4	400	2908.17	5525.95	91.29	92.48	90.12	0.91
6	600	1080.59	3173.63	91.72	91.81	91.62	0.92
8	800	1072.39	2719.83	91.57	91.86	91.28	0.92
10	1000	986.32	2286.70	91.78	91.46	92.11	0.92
12	1200	1271.00	2239.70	91.80	91.72	91.88	0.92
14	1400	1073.10	1975.48	92.14	92.21	92.08	0.92
16	1600	1179.27	1861.34	92.13	92.21	92.05	0.92
18	1800	1286.57	1737.02	92.20	92.46	91.94	0.92
20	2000	1256.00	1679.10	91.88	92.20	91.57	0.92
22	2200	2799.00	1665.23	92.11	92.42	91.79	0.92
24	2400	1477.03	1563.95	91.69	92.24	91.14	0.92
26	2600	1370.75	1438.28	91.16	91.10	91.23	0.91
28	2800	1431.13	1408.13	92.13	92.40	91.85	0.92
30	3000	1581.93	1375.01	91.10	91.13	91.06	0.91
32	3200	1655.04	1308.76	92.23	92.61	91.85	0.92
34	3400	1754.21	1275.77	91.73	91.66	91.79	0.92
36	3600	2110.51	1275.87	91.94	92.28	91.60	0.92
38	3800	1987.58	1179.22	91.89	92.05	91.74	0.92
40	4000	2168.41	1206.47	92.15	92.70	91.60	0.92
42	4200	2679.70	1259.98	91.94	92.52	91.37	0.92
44	4400	2470.30	1073.06	92.25	92.42	92.08	0.92
46	4600	2524.28	1221.51	90.92	90.26	91.60	0.91
48	4800	2198.17	1021.80	91.29	91.65	90.94	0.91
50	5000	2796.22	1153.84	91.30	91.57	91.03	0.91
52	5200	2788.94	1100.90	91.81	91.77	91.85	0.92
54	5400	2897.94	1060.17	91.54	91.85	91.23	0.92
56	5600	3233.16	1084.91	92.02	92.34	91.71	0.92
58	5800	3357.76	1041.56	91.38	91.13	91.62	0.91
60	6000	3040.89	1042.37	91.36	92.02	90.72	0.91

Various metrics, including LOSS TOK2VEC, LOSS NER, ENTS_F, ENTS_P, ENTS_R, and SCORE, shown in Table 1, were used to assess the model's performance.

The LOSS TOK2VEC metric measures the loss incurred during the training of the word vector representation component of the model. Contrary to expectations, the results indicate an increase in LOSS TOK2VEC as the number of epochs and instances increases. This suggests that the model may not effectively capture words' semantic meaning over time. For example, at E=2, #=200, the LOSS TOK2VEC is 4099.11, which increases to 3040.89 at E=60, #=6000 (Table 1).

Similarly, the LOSS NER metric, which represents the loss during the training of the NER component, exhibits a significant increase from 178.82 at E=0, #=0 to 1042.37 at E=60, #=6000. This further indicates a decrease in the model's performance over time.

On the other hand, the ENTS_F score, which is the F1 score for the named entities recognised by the model, demonstrates improvement over time. Starting from 0.00 at E=0, #=0, it reaches 91.36 at E=60, #=6000, indicating an enhancement in the model's ability to accurately identify named entities.

As shown in Table 1, the precision (ENTS_P) and recall (ENTS_R) of the model also improve as the training progresses. The precision increases from 0.00 at E=0, #=0 to 92.02 at E=60, #=6000, while the recall improves from 0.00 to 90.72 during the same period. These improvements suggest a more efficient model for recognising named entities.

The overall SCORE, which evaluates the model's performance based on various factors such as entity recognition accuracy, precision, and recall, shows improvement from 0.00 at E=0, #=0 to 0.91 at E=60,

#=6000 as shown in Table 1. This indicates the model's growing effectiveness in entity recognition as it is trained on more data.

It is important to note that the model's performance fluctuations can occur during training, as observed at E=26, #=2600, where the SCORE reduces from 0.92 to 0.91 compared to E=24, #=2400 (Table 1). Such fluctuations are common in machine learning models, likely attributed to the dataset's complexity and variability of named entities.

While the reported loss values are relatively high, it is crucial to consider the high F1 score, which indicates the model's overall accuracy in identifying named entities. It is important to interpret loss values in conjunction with performance measures, as loss values alone do not provide a comprehensive assessment of the model's effectiveness. Additionally, providing context or comparing the results with other models or baseline scores would enhance the evaluation of the model's performance.

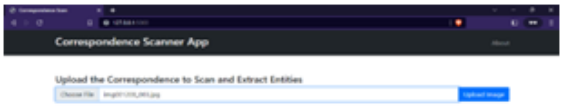

Figure 1: App Homepage

Future work could investigate the reasons for the high loss values and explore strategies to reduce them while potentially maintaining a high F1 score. Understanding the trade-off between loss metrics and performance measures can contribute to refining the model's training process and improving its effectiveness in named entity recognition.

In summary, the results demonstrate the successful training of the NER model on the dataset, showing improvement over time across various performance metrics. This suggests that leveraging SpaCy's NER capabilities and robust methodology can effectively digitise and manage data in an academic setting like Kaduna Polytechnic.

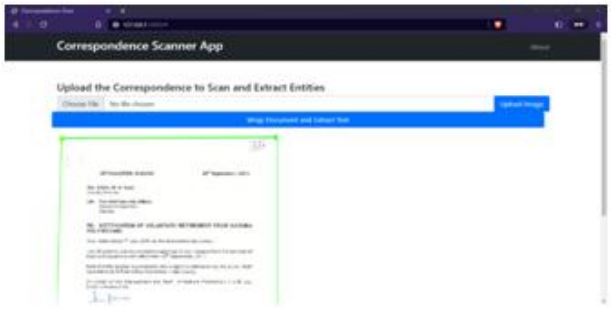

Figure 2: Uploaded Image

Figure 3: Extracted Entities

5.Conclusion

This research highlights the successful application of Named Entity Recognition (NER) using the SpaCy library for data extraction from scanned official correspondences at Kaduna Polytechnic. The model improved steadily over the training period, as various evaluation metrics showed. Its successful implementation, achieving an F1 score of 92.23%, confirmed that NER using SpaCy could effectively handle data management tasks in academic settings.

The research also contributes to the broader Natural Language Processing (NLP) field, illustrating NER's applicability in a unique context. Despite its achievements, the study recognises that accuracy and reliability will increase with more data and model refinement, suggesting areas for future work. The study concludes that NER technologies, particularly with SpaCy, hold significant potential for improving data management tasks in academic institutions and present exciting opportunities for further research and applications in the field of NLP.

Acknowledgment

The authors wish to acknowledge the funding provided for this research work by Tertiary Education Trust Fund (TETFund) under the Institutional Based Research Fund (IBR) Grants.

References

[1] Al-Qurishi, M. A., & Souissi, M. (2021). A simple and effective model for Arabic named entity recognition. Journal of Information Science, 47 (5), 634-650.
[2] Al-Rfou, R., Choe, D., Constant, N., Güngör, T., & Jones, L. (2019). Character-level language models for multilingual named entity recognition. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp.3397-3407).
[3] Al-Smadi, M., Al-Ayyoub, M., Jararweh, Y., Gupta, B. B., & Al-Betar, M. (2020). Named entity recognition for Standard Arabic using deep learning. Future Generation Computer Systems, 108, 805-818.
[4] Artetxe, M., Labaka, G., & Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp.789-798).
[5] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., . . . & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv: 2005.14165.
[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp.4171-4186).

[7] Freire, A., Santos, J. R., & Paiva, A. C. (2012). Conditional random fields for named entity recognition in Portuguese. Journal of the Brazilian Computer Society, 18 (3), 199-208.
[8] Gridach, M. (2016). Deep neural networks for named entity recognition in Arabic. In Proceedings of the 5th International Conference on Systems and Control (ICSC) (pp.1-6).
[9] Kim, Y., Jernite, Y., Sontag, D., & Rush, A. M. (2017). Structured attention networks. In Proceedings of the 5th International Conference on Learning Representations (ICLR).
[10] Kurata, G., Zhang, Y., & Matsumoto, Y. (2016). Leveraging sentence-level information with encoder LSTM for semantic role labeling. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp.1346-1351).
[11] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (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 (pp.260-270).
[12] Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30 (1), 3-26.
[13] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings.