A SPAN-BASED NAMED-ENTITY RECOGNITION METHOD FOR THESES INFORMATION-EXTRACTION ON COMPUTER SCIENCE STUDIES USING OCR TESSERACT

Undergraduate Thesis
Submitted to the Faculty of the
Department of Computer Studies
Cavite State University – Imus Campus
City of Imus, Cavite

In partial fulfillment
of the requirements for the degree
Bachelor of Science in Computer Science

RAFAELLA R. BAÑEZ
AALIHYA M. RIVERO
RYAN CHRISTIAN M. ROBLES
January 2025

TABLE OF CONTENTS

INTRODUCTION

Background of the Study

The exponential growth of academic research and digital documents has highlighted the pressing need for effective tools for managing, organizing, and extracting important data. In the computer science field, theses and technical papers are rich sources of knowledge but are often stored in formats that hinder automated processing. Named-Entity Recognition (NER) is a fundamental part in Natural Language Processing (NLP) that involves identifying and categorizing entities such as names, locations, dates, and other structured data from unstructured text (Li et al., 2019). Traditional rule-based NER methods rely on handcrafted dictionaries, keyword lists, and pattern-matching techniques to extract entities. While these methods can work well in controlled environments, they struggle with ambiguous entity names, new terms, and complex sentence structures.

Technologies such as optical character recognition (OCR) allows for the digitization of printed text from scanned documents or images. When NER is combined with OCR systems like Tesseract, it becomes possible to transform unstructured data, such as scanned documents, into a structured format that is easier to process, search, and use for various applications.

This study aims to address the difficulties involved in processing various document formats and layouts in computer science theses by developing a span-based NER technique integrated with OCR Tesseract. By integrating these technologies, the research aims to enhance the automation of thesis information extraction, supporting academic research and simplifying data organization through accurate and efficient retrieval.

Statement of the Problem

This study aims to answer "How can the researchers develop a span-based named-entity recognition (NER) method for thesis information extraction on Computer Science studies using OCR Tesseract?" Specifically, this study seeks to address the following problems:

Accurate text extraction is a fundamental challenge in processing scanned academic documents, especially when dealing with varying document conditions. "*How can OCR Tesseract be enhanced to improve the accuracy and efficiency of text extraction from scanned documents, particularly addressing challenges such as poor image quality, complex layouts, and inconsistent formatting?*

The lack of domain-specific NER models for Computer Science thesis data complicates the structured extraction of valuable information. "*What techniques can be employed to design and implement a span-based NER method tailored to extracting thesis-related information (e.g., titles, authors, keywords) specific to Computer Science studies?"*

Manual extraction of relevant thesis details is time-consuming and prone to human error, necessitating an automated solution to streamline the process. "*How can the proposed system improve the automation of thesis information extraction to support academic research and data organization?"*

Objective of the Study

This study aims to develop a span-based named-entity recognition (NER) method for thesis information extraction on Computer Science studies using OCR Tesseract

Specifically, the study aims to**:**

1. Utilize OCR Tesseract for accurate and efficient text extraction from scanned documents of Computer Science theses, focusing on handling diverse document formats and layouts.
2. Develop and implement a span-based NER method optimized for extracting key thesis-related information, such as titles, authors, keywords, and publication dates, tailored to the Computer Science domain.
3. Enhance the automation of thesis information extraction by creating a system that supports academic research and simplifies data organization through accurate and efficient information retrieval.

Time and Place of the Study

The study entitled “Developing a Span-Based Named-Entity Recognition Method for Theses Information-Extraction on Computer Science Studies using OCR Tesseract” was proposed by the researchers in Cavite State University Imus campus, which began in December 2024 and the title proposal will be conducted on January 2025 this study is under the guidance of the Undergraduate Thesis professor, Mr. Ramil Huele.

Scope and Limitations of the Study

User Interface Module. The module provides clear navigation and interactive features that guide users through tasks such as customizing search, initiating text extraction, and reviewing extracted information. It also integrates feedback mechanisms to alert users about processing status, errors, or successful operations, ensuring a smooth and reliable workflow.

Optical Character Recognition (OCR) Module. This module extracts text from scanned images of computer science theses using Tesseract OCR, ensuring accurate text recognition. By supporting a range of document formats, including PDF, DOCX, and images, the OCR Module enables flexible input handling, making it a versatile tool for digitizing thesis documents.

Named-Entity Recognition (NER) Module. This module implements a span-based NER method to identify and extract thesis-related information such as titles, authors, keywords, and publication dates, tailored to computer science studies. The NER Module is designed to handle overlapping and nested entities, ensuring comprehensive and precise extraction of relevant data. This functionality streamlines the process of organizing and utilizing academic research, significantly enhancing the accessibility and utility of extracted information.

File Management Module. This module ensures that all digitized content, including thesis titles, authors, keywords, and publication dates, is saved systematically in a structured format for future access. Users can retrieve previously processed files effortlessly, thanks to the module's user-friendly interface. This functionality supports efficient data organization by categorizing and maintaining a repository of extracted information, making it accessible for reference, further analysis, or sharing.

Customized Search. This module allows users to search and enhances usability by enabling efficient retrieval of specific details, such as a thesis title, author

name, or publication date, from a large volume of processed documents. The search mechanism is designed to handle domain-specific queries, ensuring accuracy and relevance in its results. This module supports academic research and data organization by reducing the time and effort needed to find specific information.

The limitations of this study are as follows:

- Accuracy may decrease with poorly captured images or low-quality documents, as OCR performance depends on image clarity.
- The application supports English text only.
- While supporting multiple formats (PDF, DOCX, and images), the study is limited to processing computer science theses and studies.
- The system will only cater to Computer Science related studies and theses from the Department of Computer Science at Cavite State University-Imus Campus.

Definition of Terms

Named-Entity Recognition (NER). Natural Language Processing (NLP) task that identifies and classifies named entities such as people, organizations, locations in text.

Natural Language Processing (NLP). A field of AI that focuses on the interaction between computers and human language, enabling machines to understand, interpret, and generate human language.

Optical Character Recognition (OCR). Optical Character Recognition technology that converts printed or handwritten text into machine-readable data by analyzing images of text.

Tesseract OCR. Define the specific OCR engine being used, its functionality, and its relevance to the study.

Span-Based Method. A technique in NLP where contiguous spans (sequences of tokens) are identified and classified as entities, rather than focusing on individual tokens.

User Interface. User Interface is the space where interactions between the user and a system occur, including visual elements like buttons and menus.

Theoretical Framework of the Study

The theoretical framework is based on a study by Heidarysafa M. et al. (2019) explored the use of Optical Character Recognition (OCR) technology to identify URLs and domains visited by users. Among various OCR engines, the researchers selected Tesseract-OCR, an open-source engine developed by Google in C and C++, due to its consistent performance, which rivals commercial alternatives like Transym OCR. The integration of Tesseract with Python was facilitated by the "pytesseract" library, enabling a fully Python-based implementation.

The study detailed Tesseract's architecture and process, which begins by converting input images into binary format using adaptive thresholding. Character outlines are extracted through connected component analysis, transformed into blobs, and organized into text lines. These lines are segmented into words based on character spacing, with adjustments made for proportional and fixed text. Words are further broken into character cells, and a two-stage word recognition process is applied. Results from the initial pass are refined by training a classifier, improving accuracy and overall performance. The process of this research is illustrated in Figure 1.

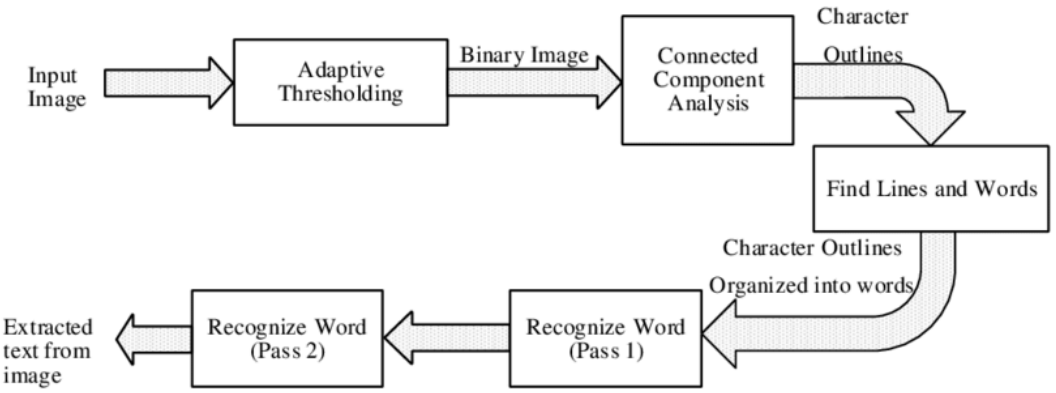

Figure 1. Theoretical Framework

Conceptual Framework of the Study

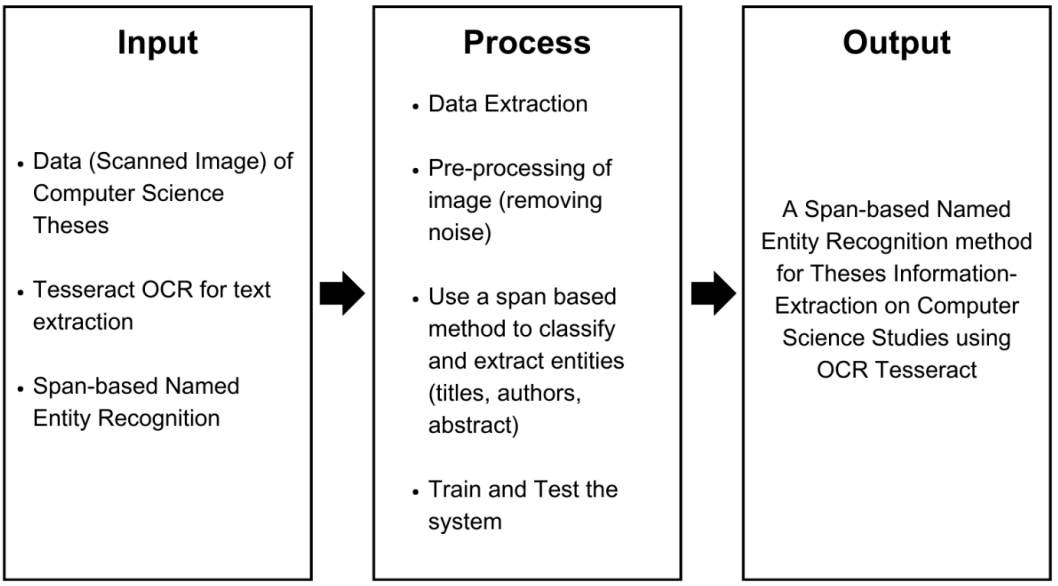

Figure 2. Conceptual Framework

REVIEW OF RELATED LITERATURE

Domain Adaptation of Named Entity Recognition for Plant Health Monitoring

This study's focus on advancing Named Entity Recognition (NER) systems for the plant health domain connects to your study, "Developing a Span-Based Named-Entity Recognition Method for Thesis Information-Extraction on Computer Science Studies Using OCR Tesseract," through its shared emphasis on improving NER performance in a specialized field. Additionally, both studies address challenges of traditional NER approaches by introducing innovations to improve adaptability. Their use of semantic entity representations parallels the study's reliance on OCR Tesseract to accurately extract and process textual data. Both approaches highlight the importance of customizing NER systems for domain-specific tasks to achieve more accurate and relevant information extraction

SpanMarker for Named Entity Recognition

The study is about SpanMarker, a machine-learning model designed for Named Entity Recognition (NER). NER is the process of finding and classifying important information (like names, dates, or organizations) in text. SpanMarker focuses on identifying these "entities" by using "spans" (specific sections of text) and assigns labels to them. It improves how efficiently and accurately these entities are identified, especially compared to older, more complex methods. The model uses advanced technologies like BERT encoders to understand the context of the text, making it better at recognizing entities. It also provides a Python library that makes it easier for researchers and developers to use the tool in real-world tasks.

The SpanMarker model achieves state-of-the-art accuracy in Named Entity Recognition (NER) tasks. It demonstrates high F1 scores (a balance of precision and recall) on several benchmark datasets, such as CoNLL03 (94.1%), OntoNotes v5.0 (91.35%), and FewNERD (70.93%). These results indicate its ability to accurately identify and classify entities across diverse datasets.

Extracting Medication Information from Typewritten Philippine Medical Prescriptions Using Optical Character Recognition (OCR) and Named Entity Recognition (NER)

The study focuses on developing a system that integrates Optical Character Recognition (OCR) and Named Entity Recognition (NER) to extract and categorize medication information from typewritten Philippine medical prescriptions. The OCR tool (Tesseract) processes prescription images into text, and an NER model (based on SpaCy) labels extracted text with relevant medical entities. With an overall F-score of 0.8816, the system shows high accuracy in recognizing and categorizing prescription information, facilitating improved data entry and medication management.

This study employs a span-based NER method for extracting structured information from documents using OCR. The use of Tesseract OCR for converting text in medical prescriptions into machine-readable formats mirrors the study approach for thesis information extraction. Insights from this study, such as OCR preprocessing and entity extraction strategies.

FINANCIAL NAMED ENTITY RECOGNITION FOR TURKISH NEWS TEXTS

It investigates the application of Named Entity Recognition (NER) techniques for identifying financial entities in Turkish news texts. It employs various deep learning models, including BERT and its variations, and introduces two newly annotated datasets to advance Turkish-language NER research. By experimenting with multilingual and language-specific models, the study achieves notable accuracy improvements in financial NER tasks.

This study works on span-based NER for thesis information extraction using OCR Tesseract. While it focuses on financial domains, the methodology of leveraging pre-trained models, annotating domain-specific datasets, and evaluating performance can provide valuable insights for extracting and categorizing computer science thesis entities.

calamanCy: A Tagalog Natural Language Processing Toolkit

The document introduces **calamanCy**, an open-source toolkit for developing NLP pipelines for Tagalog. Built on spaCy, it supports tasks such as dependency parsing, part-of-speech tagging, and named entity recognition (NER). The study provides a comprehensive evaluation of the toolkit's performance on various Tagalog benchmarks. The NER component achieves an F1-score of 90.34% using its transformer-based pipeline, showcasing competitive accuracy for Tagalog NLP tasks.

It is connected to the study of the researchers as it highlights the creation and use of a span-based Named Entity Recognition (NER) method for a low-resource language.

Table 1. Table of Comparison

	EMITPMP-OCRNER	**FNERTNT**	**SNER**	**CTNLPT**	**DANER-PHM**	**DSBNERM-TIECSSOCRT**
Year	2022	2022	2023	2023	2024	2025
Features:						
Supports Multiple Document Formats (PDF, DOCX, Images)	✖	✔	✖	✖	✖	✔
Entity-Type Coverage	✔	✔	✔	✔	✔	✔
Integration with OCR (e.g., Tesseract)	✖	✔	✖	✖	✖	✔
Automated Data Categorization	✔	✖	✔	✔	✔	✔
Customized Search	✖	✖	✖	✖	✖	✔

LEGEND:

EMITPMP-OCRNER- Extracting Medication Information from Typewritten Philippine Medical Prescriptions Using Optical Character Recognition (OCR) and Named Entity Recognition (NER) EMITPMPOCRNER

FNERTNT- Financial named entity recognition for Turkish news texts

SNER- SpanMarker for Named Entity Recognition

CTNLPT- calamanCy: A Tagalog Natural Language Processing Toolkit

DANER-PHM- Domain Adaptation of Named Entity Recognition for Plant Health Monitoring

DSBNERM-TIECSSOCRT- Developing a Span-Based Named-Entity Recognition Method for Theses Information-Extraction on Computer Science Studies using OCR Tesseract

METHODOLOGY

Materials

In order to develop the system, the following materials and technologies will be used:

Hardware. The study will use a personal computer with 8 GB RAM and AMD Ryzen 5 PRO 4650G with Radeon Graphics 3.70 GHz as its processor. For OCR scanning, the authors will use a mobile phone with an OCR-compatible app such as Google Lens for quick text extraction. The authors will use a 12 MP or above resolution camera and does have Android 5.0 or later as its operating system.

Programming Languages. The authors will be using Python as its primary language for development due to its compatibility with Tesseract OCR, NLP libraries like SpaCy, and machine learning frameworks such as TensorFlow or PyTorch

Database. The study will use SQLite for quick prototyping and small-scale applications, while MySQL/PostgreSQL are chosen for managing large-scale structured data, ensuring scalability and efficiency. MongoDB will be considered for storing unstructured data, like scanned document images or PDFs, offering flexible storage and retrieval.

Datasets. The primary data source for OCR text extraction and NER training comes from a CvSU Imus Corpus of Theses, including digital and scanned theses from repositories such as CvSU Imus Library and CvSU Imus Department of Computer Studies, while the Annotated Dataset contains manually tagged entities like author names, titles, and institutions for training the NER model.

Libraries and Tools. For text extraction, Tesseract OCR will be used, while spaCy, Hugging Face Transformers, and frameworks like PyTorch and TensorFlow are

employed to build and train the span-based NER model. Pandas will be used for efficient data handling, and Flask/Django will be considered to be used for user interface.

Plan Experimental

The system will be developed following a streamlined Software Development Life Cycle (SDLC). The first phase, Requirements Gathering and Analysis, will identify specific needs for OCR and NER integration in thesis data processing. Next, the System Design phase will focus on creating a modular architecture that includes OCR, NER, user interface, and database management modules. During the Development phase, the OCR module will be implemented first to ensure accurate text extraction from scanned theses, followed by the development and training of a span-based NER model optimized for extracting thesis-related information. In the Testing phase, both unit tests and system integration tests will be conducted using a dataset of scanned theses to evaluate performance and accuracy. Once the system passes these tests, the Deployment phase will introduce the system to end-users for feedback and refinement. Finally, Maintenance and Documentation will ensure ongoing usability, system updates, and proper user training.

Experimental Units

The system will utilize OCR Tesseract to convert scanned documents into machine-readable text, ensuring that even non-digital or image-based theses can be processed efffectively. Preprocessing steps will include text normalization to correct OCR-generated errors, as well as sentence splitting and tokenization to prepare the text for further analysis. A span-based Named-Entity Recognition (NER) model,

fine-tuned using pre-trained language model BERT, will be implemented to identify and extract entities such as thesis titles, authors, keywords, and publication date.

The system's performance will be evaluated using metrics precision, recall, and F1 score.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where:

- **Precision** (P) measures the proportion of correctly identified entities out of all entities predicted as relevant:

$$P = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

- **Recall** (R) measures the proportion of correctly identified entities out of all actual relevant entities:

$$R = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

The researchers will conduct initial testing of the system, while faculty and students will be invited to perform real-world tests and provide feedback. This collaborative testing phase will assess the system's accuracy and usability in extracting thesis-related information such as titles, authors, keywords, and publication dates.

REFERENCES:

Borovikova, M. (2024, December 13). Domain adaptation of named entity recognition for plant health monitoring. Retrieved from https://theses.hal.science/tel-04877187/

Doğru, A. H., & Karagöz, P. (2022, July 26). Financial named entity recognition for Turkish news texts. Retrieved from https://open.metu.edu.tr/handle/11511/98587

Extracting Medication Information from Typewritten Philippine Medical Prescriptions Using Optical Character Recognition (OCR) and Named Entity Recognition (NER). (2022). dlsu.edu.ph. Retrieved from https://www.dlsu.edu.ph/wp-content/uploads/pdf/conferences/research-congress-proceedings/2022/HCT-09.pdf

From Videos to URLs: A Multi-Browser Guide to Extract User's Behavior with Optical Character Recognition. (2020). Retrieved from https://www.researchgate.net/figure/Architechture-of-Tesseract-OCR_fig3_332623086

Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., & Li, J. (2019, November 7). *Dice loss for data-imbalanced NLP tasks*. arXiv.org. https://arxiv.org/abs/1911.02855

Miranda, L. J., V. (2023, November 13). CalamanCY: A Tagalog natural language processing toolkit. Retrieved from https://arxiv.org/abs/2311.07171

SpanMarker for Named Entity Recognition. (2022). Retrieved from https://F:/thesis%20(1).pdf