# Report on Deep Learning Models for Named Entity Recognition

## Introduction

This report details the implementation and results of the deep learning models developed for the Named Entity Recognition (NER) task using the CoNLL-2003 corpus. The assignment involved building, training, and evaluating models with specific architectures and using GloVe word embeddings to enhance performance.

## Task 1: Simple Bidirectional LSTM Model

### Model Description

Architecture: Embedding → BLSTM → Linear → ELU → Classifier
Embedding Dimension: 100
Number of LSTM Layers: 1
LSTM Hidden Dimension: 256
LSTM Dropout: 0.33
Linear Output Dimension: 128
Hyperparameters Tuning
Batch Size: 32
Learning Rate: 0.1
Learning Rate Scheduling: Reduced LR on Plateau factor = 0.1, patience = 4

### Performance on Development Data

Precision: 75.96%
Recall: 66.85%
F1 Score: 71.11

## Task 2: Using GloVe Word Embeddings

#### Approach to Case Sensitivity

To address the conflict between the case-insensitive nature of GloVe embeddings and the case-sensitive requirements of the NER model, we developed a strategy to extend the pre-trained word vectors with additional case information. This method involves appending a case-specific vector to the original GloVe vector, thereby preserving the rich semantic information of the embeddings while incorporating crucial case-sensitive cues that are vital for the NER task.

The function extend_vector_with_case enhances the word vectors by appending a four-dimensional case vector, indicating whether the word is uppercase, lowercase, title case, or a mix of cases. This extension enables the model to differentiate words based on their capitalization patterns, which is significant for named entity recognition as capitalization often signals the presence of named entities.

The case vector is defined as follows:

1. [1, 0, 0, 0] for words that are entirely uppercase, indicating a strong signal for named entities, especially for acronyms or initialisms.
2. [0, 1, 0, 0] for lowercase words, which are common in regular text but less so in proper names.
3. [0, 0, 1, 0] for title case words, which are often used for names, titles, or other proper nouns.
4. [0, 0, 0, 1] for words that don't fit the above categories, potentially signalling mixed case usage or special entities.

By concatenating the original GloVe vector with this case vector, the model can leverage both the semantic richness of the GloVe embeddings and the informative cues provided by the word's capitalization, enhancing its ability to recognize named entities accurately.

### Performance on Development Data

Precision: 88.74%
Recall: 90.96%
F1 Score: 89.84

## Conclusion

In this report, we've outlined the creation and assessment of deep learning models for Named Entity Recognition (NER), leveraging the CoNLL-2003 dataset. Our approach involved crafting models with specific architectures and incorporating GloVe word embeddings to boost accuracy. Initially, we deployed a straightforward Bidirectional LSTM model, achieving commendable results that validated the LSTM's capability to grasp contextual dependencies crucial for NER. Subsequently, we enhanced the model by integrating case-sensitive information into the inherently case-insensitive GloVe embeddings. This adjustment significantly improved the model's precision in identifying named entities, as reflected in the enhanced performance metrics. The strategic extension of GloVe vectors with case information has proven to be a pivotal advancement, striking a balance between semantic richness and contextual sensitivity. The outcomes highlight the efficacy of combining advanced neural architectures with enriched embeddings, pointing towards promising directions for future NER research. This project not only showcases the power of neural models in processing complex language data but also sets a precedent for innovative solutions in the evolving field of natural language processing.