A REPORT

ON

CS F429 Natural Language Processing: Assignment 1

by

Name of the student	ID number
Pavas Garg	2021A7PS2587H
Atharva Dashora	2021A7PS0127H

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
(Hyderabad Campus)

September, 2024

CONTENTS

List of Tables

List of Figures

List of Algorithms

1 Introduction

Open Information Extraction (OpenIE) is a crucial task in natural language processing (NLP) that involves extracting tuples of structured information from unstructured text. These tuples are typically in the form ⟨subject, relation, object, time, location⟩, with the goal of capturing the key relationships and entities within a sentence. OpenIE plays a fundamental role in various downstream applications, such as knowledge graph construction, question answering, and text summarization. Unlike traditional information extraction, which relies on predefined relations, OpenIE extracts relationships in an open-domain manner, making it applicable across a wide range of texts.

This task presents a supervised learning approach to OpenIE, where the goal is to design a model that extracts multiple relation tuples from a sentence. Sentences may yield more than one extraction depending on the presence of multiple relationships or entities. For example, the sentence "Burnham died of heart failure at the age of 86, on September 1, 1947, at his home in Santa Barbara, California" generates multiple tuples that include location and time information.

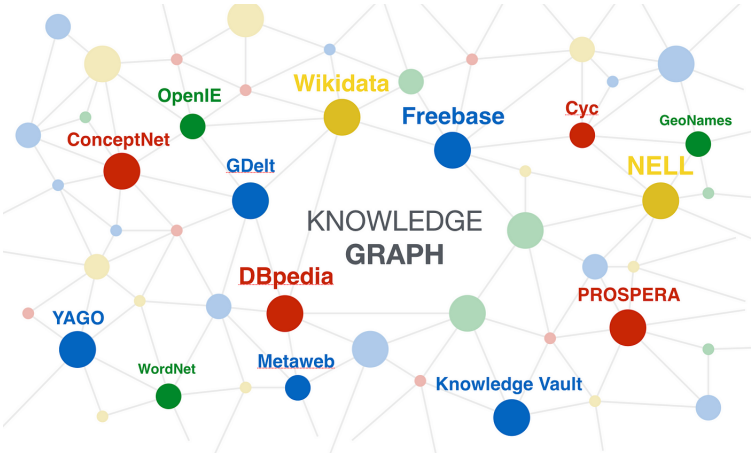

Figure 1: Knowledge Graph

This project also emphasizes error analysis to identify strengths and weaknesses of the model. Drawing inspiration from the REVERB paper, error analysis focuses on understanding where the model performs well and where it fails, providing insights for further improvements. Additionally, two variants of Word2Vec—Skip Gram and CBOW—are implemented from scratch to analyze their efficiency and quality of word embeddings.

2 Concepts

2.1 Word Embeddings

We have used BERT (Bidirectional Encoder Representations from Transformers) for generating embeddings for openIE task as it is a powerful transformer-based model that revolutionized natural language processing by providing deep contextualized word embeddings. Unlike traditional embeddings like Word2Vec or GloVe, which generate a single vector for each word regardless of context, BERT captures the meaning of a word in relation to its surrounding words. This bidirectional approach means BERT processes the entire sentence from both left-to-right and right-to-left, allowing it to generate more nuanced embeddings that reflect the word's meaning in different contexts.

For instance, the word "bank" will have different embeddings in "river bank" versus "financial bank" because BERT considers the context when generating the representation. These contextual embeddings can be directly used for downstream NLP tasks like Open Information Extraction (OpenIE), where each token's embedding can be used to classify it as a subject, relation, object, or other entity.

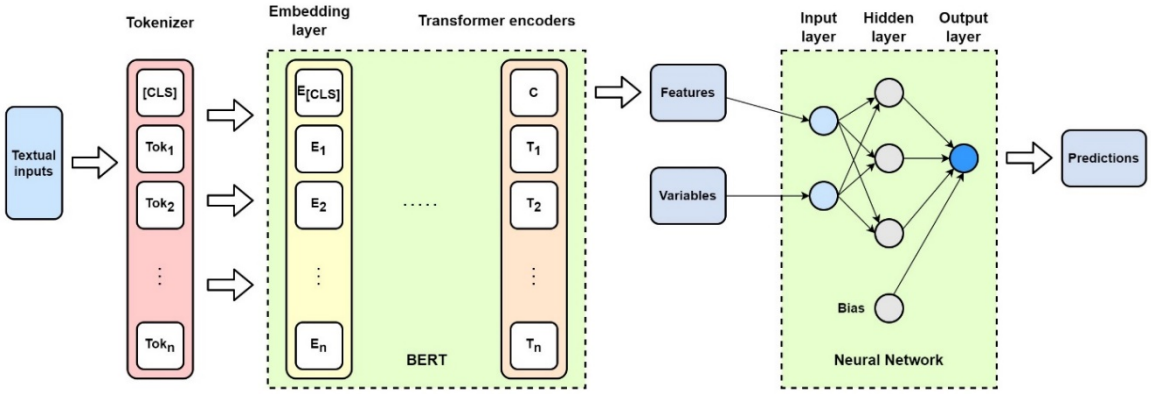

Figure 2: BERT Model

BERT embeddings are often fine-tuned for specific tasks, and in OpenIE, BERT can be used to generate token-level embeddings, which are then passed through models like BiLSTM-CRF to label tokens appropriately. This results in more accurate and context-sensitive extractions.

2.2 BiLSTM Model

A BiLSTM (Bidirectional Long Short-Term Memory) model is an extension of the LSTM architecture designed to capture dependencies in both directions within sequential data. LSTMs, by design, are capable of learning long-range dependencies by using a memory cell and gating mechanisms (input, output, and forget gates) to control the flow of information, making them highly effective in handling sequential data with varying time dependencies, such as text or speech.

In a BiLSTM, two LSTM layers are used: one processes the sequence from left to right

(forward pass), and the other processes it from right to left (backward pass). The outputs of both the forward and backward layers are concatenated at each time step, allowing the model to consider both past and future contexts simultaneously. This bidirectional structure makes BiLSTMs particularly useful for tasks like named entity recognition (NER), sentiment analysis, and relationship extraction, where understanding both preceding and following words enhances prediction accuracy.

For example, in an Open Information Extraction task, a BiLSTM model can leverage the bidirectional flow to better capture the relationship between a subject and object in a sentence, regardless of their positions. The architecture consists of an input layer (usually embeddings like BERT), the BiLSTM layers, and often a final layer (like CRF) for sequence labeling or classification. This combined architecture excels at handling complex dependencies within the data, improving the model's ability to generate accurate predictions.

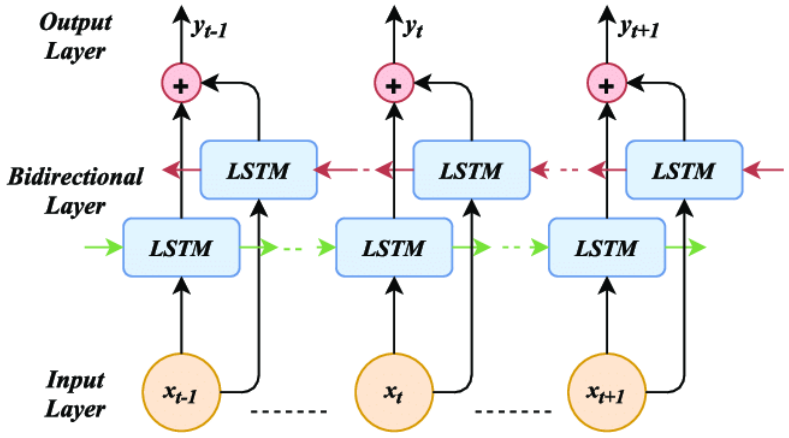

Figure 3: BiLSTM Architecture

2.2.1 BiLSTM CRF

A Conditional Random Field (CRF) layer is often added on top of a BiLSTM model in sequence labeling tasks to improve the overall prediction by considering the dependencies between output labels. While the BiLSTM layer captures context from both directions of the input sequence, the CRF layer helps ensure that the predicted sequence of labels is globally optimal by considering relationships between neighboring labels.

CRFs are particularly useful for structured prediction tasks, like named entity recognition (NER) or relationship extraction, where the prediction of one label depends on the others. For example, in a sentence, if one word is tagged as the start of an entity (e.g., 'ARG1'), it's likely that the following words are also part of the same entity, and CRF can enforce such constraints. By learning the transition probabilities between labels, the CRF layer helps ensure that the model produces consistent and valid label sequences.

In a BiLSTM-CRF model, the BiLSTM generates hidden states or embeddings for each token in a sentence, while the CRF layer takes these hidden states and predicts the

most probable sequence of labels by maximizing the conditional likelihood of the entire sequence, leading to more accurate and coherent predictions.

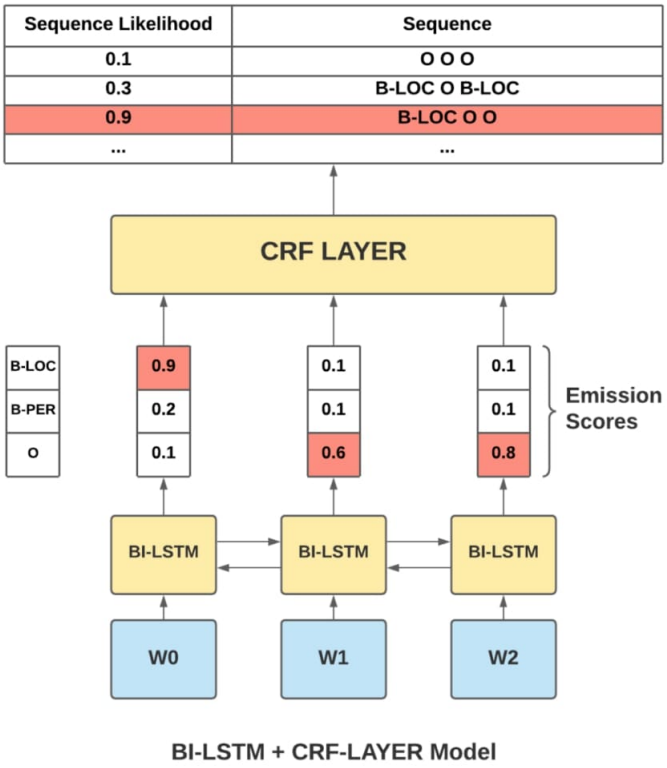

Figure 4: CRF

2.3 Word2Vec

Word2Vec is a popular technique for generating word embeddings, which are dense vector representations of words in a continuous vector space. Developed by Google in 2013, Word2Vec aims to capture semantic and syntactic relationships between words by representing words that appear in similar contexts with similar vectors. It achieves this by training on a large corpus of text, where words with similar meanings end up having close vector representations in the embedding space.

Word2Vec comes in two primary models: Skip-Gram and Continuous Bag of Words (CBOW). In the Skip-Gram model, the task is to predict the surrounding context words given a target word, while in CBOW, the goal is to predict the target word based on its surrounding words. Both models use a neural network with a single hidden layer.

These word embeddings can be used in various downstream tasks such as machine translation, document similarity, and named entity recognition, as they help encode word meaning in a way that captures both semantic similarity and contextual relationships.

2.3.1 Skip Gram

In Skip-Gram, the goal is to predict the surrounding context words (neighbors) for a given target word. For example, given the sentence "The cat sat on the mat," if the target word is "sat," the model would try to predict the words "The," "cat," "on," and "the" as its context.

Skip-Gram works by maximizing the probability of correctly predicting the surrounding words within a certain window size for each target word. A sliding window is moved across the text, and for each word (target), the model attempts to predict nearby words. The training objective is to adjust the word vectors in such a way that words occurring in similar contexts are close to each other in the vector space.

An efficient training approach for Skip-Gram involves negative sampling, which reduces computational complexity by sampling a few incorrect words ("negative samples") for each context prediction instead of updating the weights for all words in the vocabulary. This technique makes Skip-Gram scalable to large datasets. Skip-Gram embeddings are known to perform well in capturing the semantic and syntactic properties of words.

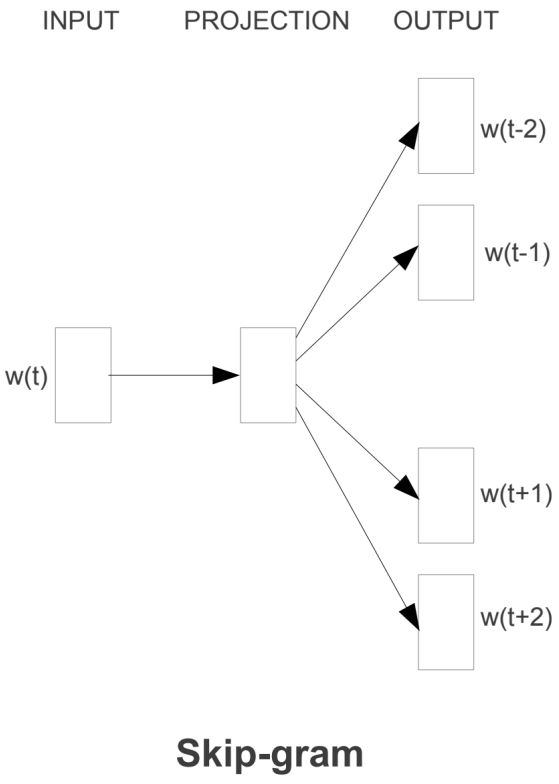

Figure 5: Skip Gram

2.3.2 CBOW

In contrast to Skip-Gram, CBOW predicts a target word based on its surrounding context words. For instance, given the sentence "The cat sat on the mat," and the context words

"The," "cat," "on," and "the," the CBOW model tries to predict the target word "sat."

CBOW works by taking the average of the vectors for the surrounding words (context) and using that to predict the vector for the target word. This method aims to maximize the likelihood of the target word appearing in the given context. Unlike Skip-Gram, which is more focused on predicting multiple context words for a single target, CBOW is designed to predict the target word from the entire context.

CBOW is generally faster to train than Skip-Gram, especially for large corpora, because it predicts one word from multiple context words, while Skip-Gram predicts many context words from a single target word.

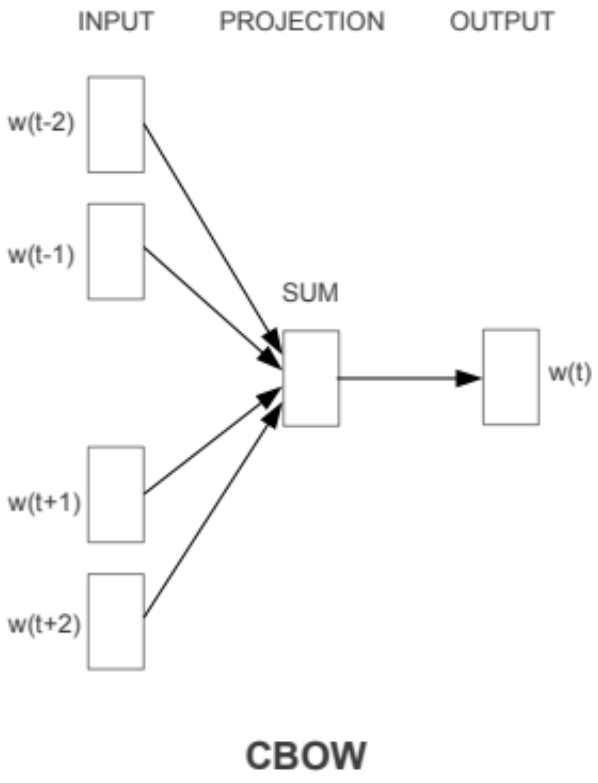

Figure 6: CBOW

3 Apparatus

Machine	Apple MacBook Pro M1 (2021)
RAM	8 GB LPDDR4X 4266MHz
Processor Name	Apple M1
Processor Specs	8 cores, 8 threads, 2.06-3.22 GHz, L1 Cache: 2MB, L2 Cache: 12MB
Python	Version 3.10

4 Dataset

The dataset for openIE task comprises of over 1e5 data points. Each token in the input sentence is labelled with one of six tags: ARG1, ARG2, REL, TIME, LOC, NONE. ARG1 and ARG2 refer to the subject and the object, respectively.

For Word2Vec task this dataset is used.

5 Methodology for OpenIE

5.1 Pre-Processing Data

For preprocessing, first the dataset was loaded from the text file using the below function then we tokenized the sentences given in the dataset using the **spacy** library, and then used the script given for combining the terms so that token and label count matches for each data point.

Algorithm 1 LOAD_DATASET

Input: $file\_path$ (Path to dataset file)
Output: $sentences$ (List of sentences), $labels$ (List of labels)
$sentences \leftarrow []$
$labels \leftarrow []$
$f \leftarrow \text{open}(file\_path, 'r')$
$lines \leftarrow f.readlines()$
$sentence \leftarrow []$
for each $line$ **in** $lines$ **do**
    $line \leftarrow line.strip()$
    if $line \neq \emptyset$ **then**
        if not $line$ **starts with** 'ARG1' **or** 'ARG2' **or** 'REL' **or** 'LOC' **or** 'TIME' **or** 'NONE' **then**
            $sentence \leftarrow line$
        else
            $current\_label \leftarrow line$
            $sentences.append(sentence)$
            $labels.append(current\_label)$
        end if
    end if
end for
return $sentences, labels$

Embeddings were generated using **BERT** and then we also added padding for each sentence to ensure uniformity across dataset as model will be trained in batches.

Tokenization: The input tokens are passed through the BERT tokenizer using the tokenizer() function. This tokenizer is special because it can split words into subword tokens if needed, which allows BERT to handle out-of-vocabulary or rare words by breaking them into smaller components. The tokenizer converts the words into the input format that BERT understands, and includes padding or truncation if necessary, though in this case, no padding or truncation is applied.

Generating BERT Embeddings: The tokenized input is fed into the BERT model to obtain contextual embeddings for each subword. BERT outputs the last_hidden_state, which contains a dense vector (embedding) for each subword token in the sequence. These embeddings capture the context and meaning of the words in their given sentence.

Mapping Subwords to Original Tokens: Since BERT's tokenizer splits some

words into multiple subword tokens, the code retrieves word_ids for each subword. This is crucial for knowing which subwords correspond to the same original word, as they need to be recombined into a single embedding.

Aggregating Subword Embeddings: Once the subword tokens are mapped to their original words using the word_ids, the next step is to aggregate the embeddings. For each original word, all its subword embeddings are averaged to create a single embedding that represents the entire word. The mean of the embeddings is used to ensure that the information from all subwords is combined effectively.

Final Embeddings: After aggregating subword embeddings for each token, the final list of embeddings for all tokens in the sentence is returned.

Algorithm 2 Get BERT Embeddings

Input: List of tokens
Output: Aggregated embeddings for each token
Initialize *inputs* ← `tokenizer`(*tokens*)
Use BERT model to compute *outputs* ← `bert_model`(*inputs*)
Extract *token_embeddings* ← `outputs.last_hidden_state`
Get word IDs *word_ids* ← `inputs.word_ids()`
Initialize empty list *aggregated_embeddings* ← []
Initialize empty list *current_token_embeddings* ← []
for each index i and *word_id* in *word_ids* **do**
  if *word_id* ≠ `None` **then**
    if *current_token_embeddings* is not empty and *word_id* ≠ *word_ids*$[i-1]$ **then**
      Compute mean of *current_token_embeddings* and append to *aggregated_embeddings*
      Reset *current_token_embeddings* ← []
    end if
    Append *token_embeddings*$[i]$ to *current_token_embeddings*
  end if
end for
if *current_token_embeddings* is not empty **then**
  Compute mean of *current_token_embeddings* and append to *aggregated_embeddings*
end if
Return *aggregated_embeddings*

For encoding the labels we used LabelEncoder(), also we added a new label called '**PADDING**' which will be used as a label for padded tokens in the sentence.

Then the dataset was split into train and validation sets using *train_test_split* function from sklearn and validation dataset size was 0.2

5.2 BiLSTM CRF Model

5.2.1 LSTM Layer

- **Input:** BERT embeddings of size 768 for each token.
- **Hidden Dimension:** 256 units. Since the LSTM is bidirectional, the hidden size becomes 512 (doubled).
- **Layers:** 2 layers of stacked LSTMs for deeper learning.
- **Bidirectional:** The LSTM captures context from both past and future directions in the sequence.

5.2.2 Dropout Layer

- A dropout probability of 0.5 is applied to the LSTM outputs to reduce overfitting.

5.2.3 Batch Normalization Layer

- The LSTM outputs are batch-normalized to speed up convergence and make the training process more stable.

5.2.4 Fully Connected Layers

- **FC1:** Linear layer with input size of 512 (from bidirectional LSTM) and output size of 256.
- **FC2:** Linear layer with input size of 256 and output size of 128.
- **FC3:** Final linear layer mapping to the number of output labels.
- **Activation:** ReLU activation function is applied after each linear transformation.

5.2.5 Conditional Random Field (CRF)

- The CRF layer models dependencies between output labels to ensure that valid label sequences are predicted.
- **Training:** The CRF computes the loss if ground truth labels are provided.
- **Inference:** The CRF decodes the most probable sequence during testing.

5.3 Training the Model

1. Hyperparameters

Learning Rate (lr = 0.001): This controls the step size during gradient descent. A smaller learning rate leads to slower convergence, but can potentially avoid overshooting minima.

Number of Epochs (num_epochs = 100): This defines how many times the model will iterate over the entire training dataset.

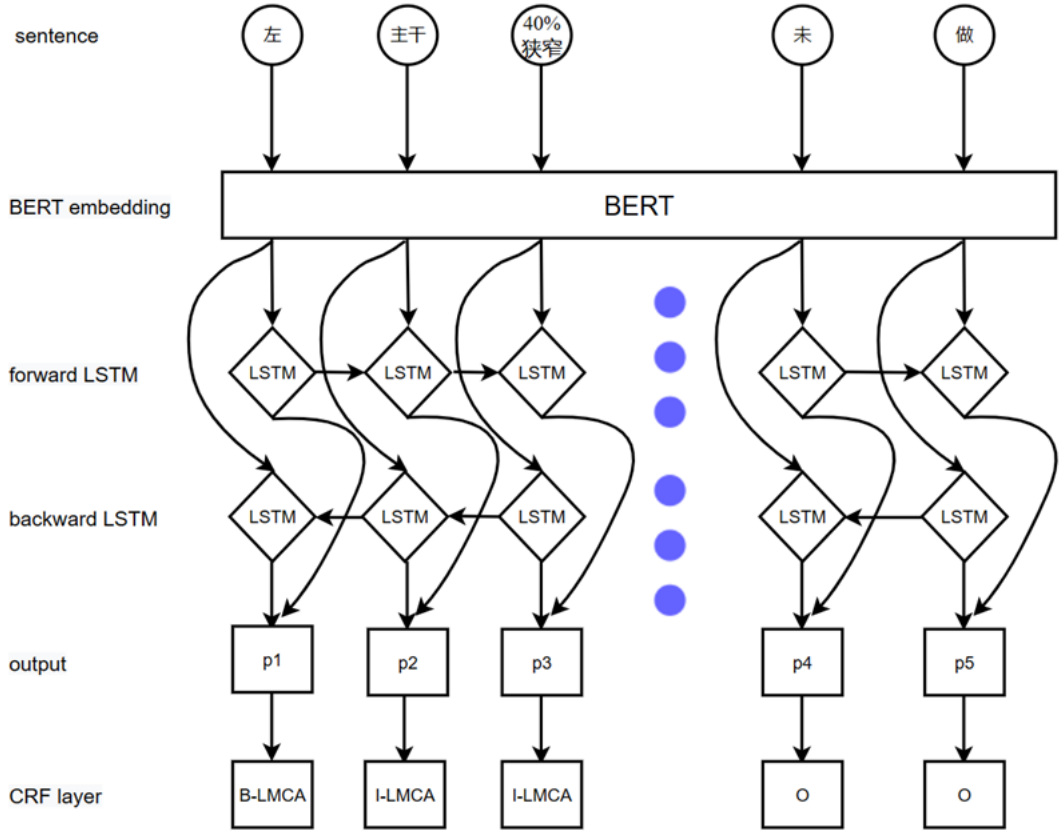

Figure 7: BERT-BiLSTM-CRF Model

Batch Size: Controlled by train_loader, the batch size determines how many samples are passed through the network before an update is made. This is important for stabilizing gradients during training.

2. Loss Function

The loss function used is **CrossEntropyLoss** with an *ignore_index* argument to skip over padding tokens:

CrossEntropyLoss: This is a commonly used loss function for multi-class classification tasks. It calculates the difference between the predicted probability distribution (logits) and the true labels.

Padding Mask: Labels corresponding to padding tokens are ignored using the *ignore_index=label_to_idx['PADDING']*. Padding is used to ensure uniform sequence lengths in batches, and this prevents these artificial tokens from influencing the loss calculation.

3. Optimizer

Adam Optimizer: Adam combines the advantages of both AdaGrad (good for sparse gradients) and RMSProp (good for non-stationary objectives). It adjusts the learning rate for each parameter dynamically, making it a popular choice for deep learning.

Algorithm 3 Training BiLSTM-CRF Model

Input: Train loader, number of epochs num_epochs, learning rate lr
Output: Trained model, loss per epoch
Initialize $optimizer \leftarrow$ `Adam(model.parameters(), lr)`
Initialize $loss\_fn \leftarrow$ `CrossEntropyLoss` with padding index ignored
Initialize $epoch\_losses \leftarrow []$
for each $epoch \in [1, num\_epochs]$ **do**
    Set model to training mode
    Initialize $total\_loss \leftarrow 0$
    for each $batch \in train\_loader$ **do**
        Get $inputs$, $labels$ from $batch$
        Create $mask$ for valid tokens where $labels \neq PADDING$
        Compute $loss \leftarrow model(inputs, tags = labels, mask = mask)$
        Zero the gradients in $optimizer$
        Backpropagate by $loss.backward()$
        Update model parameters using $optimizer.step()$
        Add $loss$ to $total\_loss$
    end for
    Compute $avg\_loss \leftarrow total\_loss/len(train\_loader)$
    Append avg_loss to $epoch\_losses$
    Print "Epoch $epoch/num\_epochs$, Loss: avg_loss"
end for

4. Masking

Masking is applied to ensure that only the valid tokens (i.e., non-padding tokens) are used to calculate the loss. The mask is computed using *labels.ne(label_to_idx['PADDING'])* condition, which checks for tokens that are not padding. This helps in preventing the model from learning patterns based on padding tokens.

5. Training Process

The model is trained over *100* epochs and for *4500* sentences, where in each epoch the model is set to training mode (*model.train()*). For each batch, the loss is computed only for valid tokens (non-padding). Gradients are backpropagated, and the optimizer updates the weights to minimize the loss. The average loss is calculated for each epoch and stored for later analysis.

5.4 Evaluation on Validation Data

The trained model was evaluated on the validation dataset giving accuracy of *61%*. This metric was found by calculating the number of labels that were matching with the ground labels and this was computed for non padding labels.

5.5 Extracting Relations

For the test dataset, pre-processing was done in a similar way by first tokenizing the sentences and then generating embeddings using BERT. The pre-trained model was used

Figure 8: Training Loss

for labelling task. Once we got the labels for each token for the test dataset, next step is to extract multiple relations from this output given by our model.

We followed different approaches:

5.5.1 Enforce a Strict Pattern on the Extractions

Make sure the extractions follow one of the possible combinations.

Standard pattern: ARG1 before REL, ARG2 after REL
Example: ARG1 (subject) is followed by REL (relation) and then ARG2 (object).
Output: "Sentence" — REL — ARG1 — ARG2 — TIME — LOC

Inverted pattern: ARG2 before REL, ARG1 after REL
Example: ARG2 (object) comes before REL (relation), followed by ARG1 (subject).
Output: "Sentence" — REL — ARG2 — ARG1 — TIME — LOC

Both arguments before REL: ARG1 before ARG2, both precede REL
Example: Both ARG1 and ARG2 occur before REL.
Output: "Sentence" — REL — ARG1 — ARG2 — TIME — LOC

Both arguments after REL: REL is followed by ARG1 and then ARG2
Example: REL is followed by ARG1 and then ARG2.
Output: "Sentence" — REL — ARG1 — ARG2 — TIME — LOC

Loosened patterns: If only ARG1 or ARG2 is present with REL, extract a partial relationship. Output includes only the available argument along with REL.

Handling Missing Elements

No relation (REL) found: If neither REL nor ARG1/ARG2 is found but time or location is detected, extract just the time or location information.

Algorithm 4 Relationship Extraction Algorithm (Strict Pattern)

Input: Model *model*, Padded embeddings *padded_embeddings*, Test data *df_test*
Output: Results *results*, Skipped count *skipped_count*
Initialize *results* ← [], *skipped_count* ← 0
for each sentence *i* in *df_test* **do**
    Get predicted labels *predicted_labels*
    Initialize flags and position markers for ARG1, ARG2, REL, TIME, LOC
    for each token in *predicted_labels* **do**
        Identify and group tokens into ARG1, ARG2, REL, TIME, LOC
    end for
    if REL is found **then**
        if Standard pattern ARG1 before REL, ARG2 after REL **then**
            Extract relation with ARG1, REL, ARG2, TIME, LOC
        else if Inverted pattern ARG2 before REL, ARG1 after REL **then**
            Extract inverted relation with ARG2, REL, ARG1, TIME, LOC
        else if Both arguments before REL or after REL **then**
            Extract relation
        end if
    else
        if Only time/location present **then**
            Extract partial information with TIME/LOC
        else
            Skip the sentence
            Increment *skipped_count*
        end if
    end if
    Append sentence results to *results*
end for
Return *results*, *skipped_count*

Output: "Sentence" — 0.5 — [empty] — [empty] — TIME — LOC

Confidence of *0.5* is assigned for these type of extractions. Otherwise confidence value of *1* is assigned for the extractions following any of the above patterns.

Complete skip:

If no meaningful relation, time, or location can be identified, the sentence is skipped, and a warning is logged.

Observations

Using this approach on test dataset, CaRB metric gave a high precision value of $\tilde{0.72}$, as most of the extractions were correct given a strict pattern was enforced, but at the same time recall value was very low $\tilde{0.212}$ indicating that many valid extractions were skipped which did not follow the above pattern, this could be verified by checking the number of generated extractions. There were $\tilde{300}$ extractions generated while test dataset contained *641* sentences. This is because many sentences were being skipped as most of the labels in them were NONE, hence not following any of the above pattern. F1 score value was *.328* and AUC was *.183*.

```
AUC: 0.183       Optimal (precision, recall, F1): [0.72  0.212 0.328]
```

5.5.2 Assign Confidence Value to Extractions

This approach follows the same methodology as the above approach, that is checking all possible patterns. The only difference in this approach is that it assigns a confidence score for each extraction.

We have two functions one for extracting confidence features and other one computes the overall confidence for a given extraction.

Algorithm 5 Extract Confidence Features

function EXTRACT_CONFIDENCE_FEATURES(sentence, rel, arg1, arg2, tokens)
    Initialize $features \leftarrow \{\}$
    $features$['covers_all_words'] $\leftarrow$ (len(arg1.split()) + len(rel.split()) + len(arg2.split()) == len(tokens))
    $last\_token\_rel \leftarrow$ last token in $rel.split()$ or empty string
    $features$['last_prep'] $\leftarrow$ boolean values for preps ['for', 'on', 'of', 'to', 'in']
    $features$['short_sentence'] $\leftarrow$ (len(tokens) ≤ 10)
    $features$['medium_sentence'] $\leftarrow$ ($10 <$ len(tokens) ≤ 20)
    $features$['long_sentence'] $\leftarrow$ (len(tokens) > 20)
    Define $wh\_words \leftarrow$ ['who', 'what', 'where', 'when', 'why', 'how']
    $features$['wh_word_left'] $\leftarrow$ True if any wh_word appears before rel
    $features$['r_matches_vw_p'] $\leftarrow$ match VW*P pattern
    $features$['r_matches_v'] $\leftarrow$ match V pattern
    $features$['sentence_starts_with_x'] $\leftarrow$ ($sentence$ starts with $arg1$)
    $features$['arg1_is_proper_noun'] $\leftarrow$ arg1.istitle()
    $features$['arg2_is_proper_noun'] $\leftarrow$ arg2.istitle()
    return $features$
end function

Algorithm 6 Confidence Score Calculation for Relationship Extraction

Sentence s, Relation r, Argument 1 $arg1$, Argument 2 $arg2$, Tokens $tokens$ Confidence Score CS
$features \leftarrow extract\_confidence\_features(s, r, arg1, arg2, tokens)$
$weights \leftarrow$ {weights for each feature}
$confidence\_score \leftarrow 0$

for each feature f in $features$ **do**

if $f \in weights$ **then** $confidence\_score \leftarrow confidence\_score + weights[f] \cdot features[f]$
$bias \leftarrow 3.0 \quad confidence\_score \leftarrow confidence\_score + bias$
$normalized\_score \leftarrow \frac{(confidence\_score - min\_score)}{(max\_score - min\_score)}$
$CS \leftarrow \max(0, \min(normalized\_score, 1))$
return CS

Features for confidence extraction were taken from the [FSE11] paper, weights for each feature were slightly modified in order to get better results.

Observations

Using this approach we got almost similar results, infact we observed a decrease in precision value resulting in a lower F1 score. AUC value was also slightly less about *.158*.

```
AUC: 0.158      Optimal (precision, recall, F1): [0.701 0.212 0.326]
```

5.6 Error Analysis

The following results have been prepared by comparing our extractions on the test dataset with the gold extractions.

Incorrect Extractions	**Percentage**
Correct relation, incorrect arguments	36.97%
Correct relation, incorrect argument order	0.00%
Correct relation, missing arguments	0.47%
N-ary relation	0.00%
Non-contiguous relation phrase	2.37%
Imperative verb	0.00%
Overspecified relation phrase	11.37%
Other (POS/chunking errors)	48.82%

Table 1: Incorrect Extractions

Most of the incorrect extractions that were classified, given by our model are due to arguments being incorrect. This is mainly because our model was only trained on $\tilde{4}500$ sentences and over 100 epochs, hence could not label the arguments correctly. There are no extractions with n-ary relations because the algorithms we used above have a strict pattern enforcing the argument count to be two. Significant chunk of extraction have overspecified relation phrase too. As there are no such extractions where argument order

is incorrect signifies that CRF layer is effective at preserving sequence order. Also there were no incorrect extractions due to imperative verbs concludes that model is effectively distinguishing between declarative and imperative sentences this could be because BERT is pre-trained on a large dataset, hence able to give appropriate embeddings.

Missed Extractions	**Percentage**
Could not identify correct arguments	100.00%
Relation filtered out by lexical constraint	0.00%
Identified a more specific relation	0.00%
POS/chunking error	0.00%

Table 2: Missed Extractions

All the extractions that were missed are because of incorrect labelling of arguments. We also observed that there were lot of sentences for which no extractions were being produced, this was due to the fact that model predicted all labels as NONE for these sentences, hence there were no valid arguments present. This could be due to model being not properly trained because of computing constraints.

5.6.1 Examples

Model was able to produce accurate extractions for the following sentences:

1. 32.7% of all households were made up of individuals and 15.7 % had someone living alone who was 65 years of age or older.

2. A CEN forms an important but small part of a Local Strategic Partnership.

Some sentences where model *predicted all labels as NONE* are given below:

1. A partial list of turbomachinery that may use one or more centrifugal compressors within the machine are listed here

2. Although under constant attack from kamikazes as well as fighters and dive-bombers , " Hazelwood " came through the invasion untouched and on the night of 25 February sank two small enemy freighters with her guns

6 Methodology for Word2Vec

6.1 Data Cleaning and Tokenization

1. **Corpus Input:** The process begins with a raw text corpus that may contain various forms of textual noise, including punctuation, numbers, and special characters. The corpus is input as a single string. In our case, there were various japanese texts and unknown tokens which had to be taken care of, and we removed roman numerals as well.

2. **Removal of Non-Alphabetic Characters:** To ensure the integrity of the data, a regular expression is employed to remove Roman numerals and any numeric characters. This helps to focus solely on the textual content, which is essential for meaningful word representation.

3. **Punctuation Elimination:** All punctuation marks are stripped from the text using a translation table. This step further cleans the text and avoids introducing any noise that could interfere with word tokenization.

4. **Tokenization:** The cleaned corpus is then tokenized into sentences by splitting it at periods. Each sentence is subsequently converted into a list of words. This hierarchical structure (sentences containing words) is crucial for further processing and analysis.

6.2 Filtering Sentences and Words

1. **Minimum Sentence Length:** To maintain a quality dataset, sentences are filtered based on a minimum length threshold. Sentences shorter than a specified number of words (e.g., 8 words) are discarded. This helps ensure that the remaining sentences contain sufficient contextual information for training.

2. **Word Frequency Analysis:** A frequency count of all words across the entire corpus is performed using the `Counter` class from the `collections` module. This step allows the identification of rare words that may not contribute meaningfully to the model.

3. **Rare Word Removal:** Sentences are further filtered to exclude words that occur less frequently than a predefined threshold (e.g., 10 occurrences). This reduces the vocabulary size and focuses on more commonly used words, which are more likely to yield useful embeddings.

4. **Final Sentence Filtering:** After filtering for rare words, any sentences that fall below the minimum sentence length threshold are removed. This final check ensures that all sentences in the dataset are suitable for training purposes.

6.3 Output

The preprocessing function returns a list of cleaned and filtered sentences, where each sentence is represented as a list of words. This structured output serves as the input for subsequent modeling tasks, such as training the CBOW and Skip-Gram models in Word2Vec.

6.4 Skip-Gram Model

The Skip-Gram model is a neural network-based approach for learning word embeddings by predicting the context words given a target word. The following sections outline the key components of the Skip-Gram model, including the architecture, training process, and methods for generating predictions.

6.4.1 Model Architecture

The Skip-Gram model consists of an input layer, a hidden layer, and an output layer:

1. **Input Layer:** The input layer consists of one-hot encoded vectors representing the target word. Each vector has a dimension equal to the vocabulary size, where only the index corresponding to the target word is activated.

2. **Hidden Layer:** This layer contains a fixed number of neurons (in this implementation, 20) and does not have an activation function. The weights connecting the input layer to the hidden layer are randomly initialized. The output from the hidden layer is a dense representation of the target word.

3. **Output Layer:** The output layer also consists of a one-hot encoded vector representing the context words. The model uses softmax activation to compute the probability distribution over the vocabulary for context words.

6.4.2 Training Process

The training process involves several key steps:

1. **Weight Initialization:** Weights for the connections between the input and hidden layers (W) and the hidden and output layers (W_1) are initialized randomly within a specified range.

2. **Loss Calculation:** The model employs the cross-entropy loss function to quantify the difference between the predicted probabilities and the actual context words. The loss is computed for each training example by aggregating contributions from all context words.

3. **Feedforward Propagation:** For each training instance, the model computes the hidden layer activations (h) by taking the dot product of the input vector with the weight matrix W. The output layer activations (u) are then calculated using the hidden layer output and the weight matrix W_1. The softmax function is applied to obtain the predicted probabilities (y) for context words.

4. **Backpropagation:** The error is computed as the difference between the predicted probabilities and the actual context word vector. Gradients are calculated for both sets of weights (W and W_1) using the chain rule, allowing for updates to be made via gradient descent.

5. **Adaptive Learning Rate:** The learning rate is adjusted dynamically based on the iteration number to improve convergence and optimize training efficiency.

6.4.3 Prediction Mechanism

To generate predictions for context words given a target word:

1. **Input Representation:** A one-hot encoded vector for the target word is created, similar to the training phase.

2. **Feedforward Calculation:** The hidden layer activation is calculated, followed by the output layer activation. The softmax function is applied to produce a probability distribution over the vocabulary.

3. **Top Context Words Retrieval:** The predicted probabilities are sorted, and the top context words are retrieved based on the highest probabilities. This provides insight into the most likely context words for the given target word.

6.4.4 Output and Embeddings

The model provides two key outputs:

1. **Embedding Matrix:** The learned word embeddings can be accessed as a matrix where each row corresponds to a word in the vocabulary.

2. **Individual Word Embedding:** The embedding for a specific word can be retrieved using its index in the vocabulary.

6.5 Skip-Gram Model with Negative Sampling

The Skip-Gram model with negative sampling is an enhancement of the traditional Skip-Gram approach, designed to improve training efficiency and the quality of learned embeddings. This method addresses the challenge of efficiently updating the weights when training on large corpora by focusing on a small number of context words and selected negative samples.

6.5.1 Model Architecture

Similar to the standard Skip-Gram model, the architecture includes an input layer, a hidden layer, and an output layer, with modifications to incorporate negative sampling.

1. **Input Layer:** The input layer remains a one-hot encoded vector for the target word, as in the original Skip-Gram model.

2. **Hidden Layer:** The hidden layer consists of a fixed number of neurons (20 in this implementation) and serves as the representation of the target word.

3. **Output Layer:** The output layer produces probabilities for both positive and negative samples. In this case, softmax activation is still used, but the focus is on a limited set of output neurons that correspond to the positive context words and a selected number of negative samples.

6.5.2 Training Process with Negative Sampling

The training process under negative sampling includes the following steps:

1. **Weight Initialization:** Weights are initialized similarly to the standard Skip-Gram model, creating matrices W and W_1.

2. **Loss Calculation:** Negative sampling alters the loss function. The model calculates the loss for positive samples (actual context words) and negative samples (randomly chosen words not in the context).

3. **Feedforward Propagation:** For each training instance, the hidden layer activations (h) are computed as before. However, instead of calculating the full softmax over the entire vocabulary, we focus on the context words and the negative samples, thus optimizing computation.

4. **Negative Sampling:** A small number of words from the vocabulary, which are not part of the context words, are randomly selected as negative samples. The number of negative samples significantly affects training efficiency and model performance.

5. **Backpropagation:** The error is calculated for both positive and negative samples. Gradients are computed, allowing for the weights W and W_1 to be updated using gradient descent.

6. **Adaptive Learning Rate:** As in the traditional model, the learning rate is adjusted throughout training to enhance convergence.

6.5.3 Choosing the Number of Negative Samples

Selecting the number of negative samples is critical for balancing training speed and the quality of the learned representations. In this implementation, we chose to use 5 negative samples based on several factors:

1. **Empirical Results:** Research and empirical studies suggest that using around 5 negative samples often strikes a good balance between computational efficiency and performance. This number allows the model to learn effectively from a meaningful contrast between positive and negative examples.

2. **Training Time vs. Quality:** Increasing the number of negative samples can improve the model's ability to distinguish between similar words, but it also significantly increases training time. Five samples have been found to provide sufficient learning without excessively prolonging the training process.

3. **Context Size Consideration:** Given the context window size of 2, five negative samples provide a reasonable number of contrasts, allowing the model to learn discriminative features effectively while keeping the computational overhead manageable.

6.5.4 Prediction Mechanism

The prediction mechanism follows similar steps as the traditional Skip-Gram model, with adaptations to incorporate negative samples:

1. **Input Representation:** A one-hot encoded vector for the target word is created.

2. **Feedforward Calculation:** The hidden layer activation and output layer probabilities are calculated, focusing only on the selected context words and negative samples.

3. **Top Context Words Retrieval:** The predicted probabilities are sorted, and the most likely context words are returned based on the computed scores.

6.6 Continuous Bag of Words (CBOW) Model

The Continuous Bag of Words (CBOW) model is an approach for learning word embeddings by predicting a target word based on its surrounding context words. Unlike the Skip-Gram model, which uses a target word to predict its context, CBOW takes multiple context words as input to predict a single target word.

6.6.1 Model Architecture

The architecture of the CBOW model includes an input layer, a hidden layer, and an output layer. The main components are as follows:

1. **Input Layer:** The input layer consists of context words represented as one-hot encoded vectors. For each target word, the model utilizes multiple context words defined by a specified context window.

2. **Hidden Layer:** The hidden layer contains a fixed number of neurons (20 in this implementation) and is responsible for learning the representation of the context words. The context vectors are summed to create a single input vector for the hidden layer.

3. **Output Layer:** The output layer consists of a softmax function that produces a probability distribution over the vocabulary, allowing the model to predict the target word based on the learned context representation.

6.6.2 Training Process

The training process for the CBOW model involves several steps:

1. **Weight Initialization:** Weights are initialized for the input-to-hidden layer and hidden-to-output layer, denoted as W and W_1, respectively.

2. **Feedforward Propagation:** For each training instance, the model computes the context vector by summing the one-hot encoded vectors of the context words. This context vector is then used to compute the hidden layer activations.

3. **Output Calculation:** The output layer computes the probabilities of each word in the vocabulary being the target word. This is achieved using the softmax function on the output layer.

4. **Loss Calculation:** The loss is calculated using the cross-entropy loss function, comparing the predicted probabilities with the actual target word's one-hot encoded vector. This loss guides the optimization process.

5. **Backpropagation:** Gradients of the loss with respect to the weights are computed, and the weights W and W_1 are updated using gradient descent.

6. **Adaptive Learning Rate:** Similar to the Skip-Gram model, the learning rate is adjusted throughout the training process to enhance convergence.

6.6.3 Prediction Mechanism

The prediction mechanism for the CBOW model follows these steps:

1. **Input Representation:** The context words are represented as one-hot encoded vectors, which are summed to form the context vector.

2. **Feedforward Calculation:** The context vector is passed through the hidden layer to obtain the hidden layer activations, which are then used to compute the output layer probabilities.

3. **Target Word Retrieval:** The predicted probabilities are processed to identify the target word with the highest probability, which is returned as the output.

6.6.4 Advantages of CBOW

The CBOW model is particularly advantageous in scenarios where:

1. **Simplicity and Efficiency:** By using context words to predict a single target word, CBOW is computationally efficient, especially in large datasets with extensive vocabularies.

2. **Contextual Learning:** The model effectively captures the semantic meaning of words by leveraging multiple context words, leading to high-quality word embeddings.

6.7 Training Methodology for Skip-Gram Model

The training of the Skip-Gram model is a systematic process aimed at learning meaningful word embeddings from a given corpus. This section details the data preparation, batch generation, training loop, and overall training duration.

6.7.1 Data Preparation

1. **Sentence Tokenization:** The corpus is tokenized into individual sentences.

2. **Vocabulary Construction:** A dictionary is created where each unique word is assigned an index.

   The context window size is defined as 2, meaning that for each target word, the model will consider two words on either side as context. The vocabulary size V is computed based on the unique words in the dataset.

6.7.2 Batch Generation

Given the large size of the dataset, generating word-context pairs for each training iteration can be computationally intensive and can consume a significant amount of memory. To address this challenge, we implement a batch generation strategy. This involves processing the data in smaller, manageable chunks, which is essential for efficient memory utilization and faster training times.

In our implementation, we selected a subset of 5000 sentences from the dataset. This selection resulted in approximately 650 batches, with each batch containing 1000 word-context pairs. The batching approach not only streamlines the training process but also ensures that the model learns effectively from a diverse range of word-context pairs.

6.7.3 Training Loop

The training of the Skip-Gram model is organized into a series of epochs. In our case, we set the number of epochs to 2. The training loop encompasses the following steps:

Algorithm 7 Training Algorithm for Skip-Gram Model

```
Input: Corpus of sentences, Vocabulary, Context Window Size, Batch Size
Initialize: Model with random weights
for each epoch in number of epochs do
    Print "Starting epoch"
    for each batch in batch generator do
        Update model training data with current batch
        model.train(mytol=1e-4, maxepochs=1)
    end for
    Print "Finished epoch"
end for=0
```

6.7.4 Training Duration

The training duration for the Skip-Gram model was observed to be approximately 23 minutes when utilizing the selected 5000 sentences. This time frame reflects the computational complexity associated with processing a substantial number of word-context pairs and optimizing the model's weights for effective embedding learning.

In a separate experiment involving the Skip-Gram model with negative sampling, we utilized a smaller subset of 1000 sentences. Despite the reduced data size, the training process proved to be more computationally intensive, resulting in a total training duration of approximately 150 minutes. This increase in time can be attributed to the additional complexity introduced by negative sampling, which requires the model to handle both positive and negative examples during training.

In conclusion, the methodology employed for training the Skip-Gram model was designed to efficiently manage large datasets while ensuring effective learning of word embeddings. The incorporation of batch generation and careful selection of training subsets were critical factors contributing to the overall success of the training process.

6.8 Mean Reciprocal Rank (MRR) Metric

The Mean Reciprocal Rank (MRR) is a statistical measure used to evaluate the effectiveness of systems that return a ranked list of items. It is particularly useful in the context of information retrieval and recommendation systems, where it assesses how well a model can predict relevant items based on a given input.

The MRR is defined as the average of the reciprocal ranks of the first relevant item in the list of results returned by the model. Formally, it can be expressed as:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$

where $|Q|$ is the total number of queries, and rank_i is the rank position of the first relevant item for the i^{th} query. An MRR value closer to 1 indicates better

performance, as it suggests that the relevant items are ranked higher in the returned list.

In the context of the Skip-Gram model, MRR serves as a useful metric for evaluating the quality of word embeddings, particularly how effectively the model captures semantic relationships between words.

6.8.1 Results for Skip-Gram Model

We evaluated the performance of the Skip-Gram model with varying window sizes. The following results were obtained for different window sizes:

- **Window Size 2:** The MRR was found to be 0.2013. This value indicates that, on average, the first relevant word appears at approximately the fifth position in the ranked list. The relatively moderate performance suggests that while the model captures some semantic relationships, there is room for improvement in the context it considers.
- **Window Size 3:** The MRR improved to 0.2083. This increase in performance indicates that the model benefited from a larger context window, allowing it to consider more surrounding words when predicting relevant items. As a result, the model was able to rank relevant words slightly higher compared to the smaller window size.
- **Window Size 5:** The MRR decreased slightly to 0.2032. This drop suggests that while a larger context can help capture relationships, there can be diminishing returns if the context becomes too broad. In this case, the model may have started to incorporate less relevant words, which negatively affected the ranking of the true positives.

6.8.2 Results for Skip-Gram with Negative Sampling

When employing the Skip-Gram model with negative sampling, we utilized a smaller dataset of only 1000 sentences. Due to this limitation, the MRR values obtained may not reflect the model's full potential, as the reduced data size can hinder the learning of robust word embeddings. Consequently, the MRR values are expected to be lower compared to those obtained from the full Skip-Gram model.

The results for various window sizes are as follows:

- **Window Size 2:** The MRR was 0.0861. This low value indicates that the model struggles significantly in ranking relevant items, likely due to insufficient training data and the added complexity of negative sampling.
- **Window Size 3:** The MRR improved slightly to 0.0972. The increase suggests that expanding the context window helped the model marginally, allowing it to consider a broader set of relevant words.
- **Window Size 5:** The MRR further improved to 0.1034. The consistent increase across window sizes indicates that even with negative sampling, a larger context contributes positively to the model's performance.

In general, negative sampling has the potential to improve performance compared to standard Skip-Gram approaches because it focuses the learning on a subset of positive and negative examples, making it computationally efficient. However, due to hardware constraints, we were limited to training on fewer data points (1000 sentences). This limitation is a key reason why the MRR values for the Skip-Gram model with negative sampling are relatively lower compared to those obtained from the Skip-Gram model trained on a larger dataset.

6.9 Training Time Considerations

The training time for the word embedding models was significantly impacted by the large vocabulary size, which consisted of approximately 10,000 unique words. This extensive vocabulary leads to increased complexity in the model's computations, particularly during the training phase of both the Skip-Gram model and the Skip-Gram model with negative sampling.

6.9.1 Factors Contributing to Long Training Time

The following factors contributed to the long training time:

(a) Increased Computational Load: With a vocabulary size of around 10,000 words, the weight matrices W and W_1 in the models are considerably large. Each training iteration involves matrix multiplications that grow with the vocabulary size, leading to higher computational demands. Specifically, each weight update requires operations proportional to $O(V \times N)$, where V is the vocabulary size and N is the number of neurons in the model.

(b) Batch Processing: Although batch processing was employed to manage memory usage, each batch still contains a substantial amount of data due to the large number of unique words. This necessitates additional iterations to process all word-context pairs effectively, thus extending the overall training time.

(c) Complexity of Training Algorithms: Both models utilize backpropagation through multiple layers, requiring gradient computations for all weights in the network. As the vocabulary increases, more computations are necessary for updating these weights, contributing to the increased training time.

(d) Convergence Requirements: Given the complexity introduced by a large vocabulary, the models may require more epochs to converge to an optimal solution. This is particularly true in scenarios where the model is sensitive to initial weight distributions and learning rates.

The total training time for the Skip-Gram model was approximately 23 minutes when using 5,000 sentences, primarily due to the significant computation needed to process each word-context pair within the batches, compounded by the large vocabulary size.

6.9.2 Results for Skip-Gram and Skip-Gram with Negative Sampling

In the case of the Skip-Gram model, the following results were observed:

- Window Size 2: Mean Reciprocal Rank (MRR) = 0.2013.
- Window Size 3: MRR = 0.2083 (indicating improved performance).
- Window Size 5: MRR = 0.2032 (suggesting diminishing returns from increasing the context window size).

The training time for the Skip-Gram model was notably long due to the factors mentioned above, which compounded the computational load significantly.

For the Skip-Gram model with negative sampling, we utilized a smaller dataset of 1,000 sentences, yielding the following MRR results:

- Window Size 2: MRR = 0.0861.
- Window Size 3: MRR = 0.0972.
- Window Size 5: MRR = 0.1034.

While negative sampling generally enhances training efficiency by focusing on a subset of positive and negative examples, the limited dataset size, combined with the inherent complexity of the model due to the large vocabulary, resulted in longer training times and lower MRR scores compared to the standard Skip-Gram approach.

Model	Learning Rate	Number of Neurons	Number of Negative Samples	Training Time
Skip-Gram	0.005	20	N/A	23 minutes
Skip-Gram with Negative Sampling	0.005	20	5	150 minutes

Table 3: Summary of Hyperparameters and Training Time for the Models

7 References

References

[FSE11] Anthony Fader, Stephen Soderland, and Oren Etzioni. "Identifying relations for open information extraction". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Edinburgh, United Kingdom: Association for Computational Linguistics, 2011, pp. 1535–1545. ISBN: 9781937284114.

8 Glossary

Open Information Extraction	Method for extracting structured information from unstructured text, typically focusing on extracting relationships in the form of triples (subject, relation, object) from natural language sentences.
Embedding	A technique to represent words or phrases as vectors in a continuous vector space, capturing semantic meanings and relationships between them.
N-ary Relation	A relation involving more than two entities or arguments, as opposed to binary relations that involve only a subject and object.
POS Tagging	The process of assigning grammatical categories (such as nouns, verbs, adjectives) to individual words in a sentence to understand their roles in the context.
Stemming	Chops off word endings, sometimes creating nonsensical words.
Lemmatization	Takes context into account, finding the true root of a word (e.g., "running" becomes "run").

Table 4: Glossary