NLP 203 Assignment 3

Ishika Kulkarni
ikulkar1@ucsc.edu

March 8, 2025

1 Problem 1

Integer Linear Program (ILP) for Sequence Labeling

We define a binary decision variable x_{ij}^t for each possible transition from tag i to tag j at timestep t:

$$x_{ij}^t = \begin{cases} 1, & \text{if transition from } i \text{ to } j \text{ occurs at time } t \\ 0, & \text{otherwise} \end{cases}$$

1. Objective Function We aim to maximize the total transition score across all time steps:

$$\max \sum_{t=0}^{T} \sum_{i,j} s_{ij}^t x_{ij}^t$$

where: - s_{ij}^t is the given score for transitioning from tag i to tag j at time t. - x_{ij}^t is the binary variable indicating whether the transition occurs.

This objective ensures that the highest-scoring sequence of transitions is selected.

2. Constraints for Exactly One Transition Per Timestep At each timestep t, exactly one transition must occur, meaning the model must pick exactly one transition from any tag i to some tag j:

$$\sum_{i,j} x_{ij}^t = 1, \quad \forall t \in \{0, \dots, T\}$$

Since x_{ij}^t is binary, this equation ensures that exactly one of the x_{ij}^t values is set to 1 at each timestep. This guarantees that each timestep has exactly one active transition.

3. Constraints for Valid Paths To enforce valid paths, we ensure that if there is a transition from tag i to tag j at timestep t, then at $t+1$, the sequence must continue with some transition from j to another tag k:

$$\sum_k x_{jk}^{t+1} \geq x_{ij}^t, \quad \forall t, i, j$$

If $x_{ij}^t = 1$, then $\sum_k x_{jk}^{t+1}$ must be at least 1, ensuring that j transitions to some valid tag k at $t+1$. This maintains continuity in the sequence, preventing the model from ”stopping” in the middle of a sequence.

4. BIO Tagging Constraint (Tag I Cannot Follow Tag O) In BIO tagging, an ”I” (Inside) tag should never immediately follow an ”O” (Outside) tag. To prevent such invalid transitions, we add the following constraint:

$$x_{O,I}^t = 0, \quad \forall t$$

This explicitly prohibits transitions where tag O is immediately followed by tag I. Ensures that the model only allows valid BIO sequences.

2 Problem 2

Decision Variables:

- $x_{ij} = 1$ if word w_i is the parent of word w_j.
- $x_{ij} = 0$ otherwise, where $i \in \{0, 1, \dots, n\}$, and $j \in \{1, \dots, n\}$.

Objective Function:

$$\text{Maximize } \sum_{i=0}^{n} \sum_{j=1}^{n} \text{score}(w_i, w_j) \cdot x_{ij}$$

Parent Constraints:

- Each word except the root has exactly one parent:

$$\sum_{i=0}^{n} x_{ij} = 1 \quad \text{for each } j \in \{1, \dots, n\}$$

- The root w_0 has no parent (implicitly handled).

Projectivity Constraints:

- If w_i is the parent of w_j, then for each word w_k between w_i and w_j, the parent of w_k must also be between w_i and w_j.
- This can be expressed as:

$$x_{ij} + \sum_{l \notin [i,j]} x_{lk} \leq 1 \quad \text{for each } i, j, k$$

Acyclicity Constraints:

- Introduce depth variables d_j, where d_j represents the depth of word w_j in the tree (distance from the root).
- For every edge (i, j) where $i \neq 0$, enforce:

$$d_i - d_j + n \cdot x_{ij} \leq n - 1 \quad \text{for each } i, j$$

- For edges from the root $(i = 0)$:

$$d_j \geq x_{0j} \quad \text{for each } j \in \{1, \dots, n\}$$

- This ensures that the parent has a smaller depth for each parent-child pair than the child, enforcing a strict hierarchy.

3 Problem 3

3.1 Introduction

Abstractive text summarization is all about creating clear and concise summaries that really capture the main ideas from a piece of writing. Unlike extractive summarization, which pulls out whole sentences directly from the text, abstractive methods actually generate new sentences that rephrase what's being said.

In this assignment, we're working on building an abstractive summarization system using a sequence-to-sequence (seq2seq) model that's enhanced with an attention mechanism. We'll be using the CNN/Daily Mail dataset for training, including news articles and their longer summaries. To ensure our summaries stay true to the original content, we'll evaluate the model's performance using ROUGE metrics, focusing on how well the key information is preserved.

3.2 Dataset

The CNN/Daily Mail dataset includes over 300,000 news articles, each with human-written highlights summarizing the content. This dataset is organized into training, development, and test sets, featuring 287,227, 13,368, and 11,490 article-summary pairs.

Such characteristics make it a great benchmark for training and evaluating models that excel at abstractive summarization.

It offers a challenging yet organized setting for developing seq2seq models with attention mechanisms, ultimately enhancing summarization quality.

3.3 Model 1 Baseline Implementation

Model 1 is a Seq2Seq model with attention, using GRU-based encoder-decoder architecture for abstractive summarization.

3.3.1 Architecture

- **Encoder:** A GRU-based encoder processes the input sequence (source text) and generates hidden states. The encoder consists of an embedding layer followed by a GRU layer.

- **Decoder:** A GRU-based decoder generates the output sequence (summary) token by token. It uses an attention mechanism to focus on relevant parts of the input sequence while decoding.

- **Attention Mechanism:** A fixed attention layer computes context vectors by combining the encoder's hidden states and the decoder's current state. This helps the decoder focus on important parts of the input sequence.

- **Output Layer:** A fully connected layer maps the decoder's hidden state to the vocabulary size, producing a probability distribution over the target vocabulary.

3.3.2 Implementation

- **Input Handling:** The input text is tokenized and converted into sequences of indices using a vocabulary.

- **Encoder:** The encoder processes the input sequence and produces hidden states.

- **Decoder:** The decoder generates the output sequence one token at a time, using the encoder's hidden states and attention.

- **Output Generation:** During inference, the model generates summaries by iteratively predicting the next token using the previously generated token as input.

- **Teacher Forcing:** During training, the ground truth is used as the next input to the decoder.

3.3.3 Hyperparameters

Hyperparameter	**Value**
Batch Size	16
Embedding Dimension	128
Hidden Dimension	128
Max Sequence Length	256
Number of Epochs	10
Learning Rate	0.001
Gradient Accumulation	4

Table 1: Hyperparameters and their values

3.3.4 Training

- The model is trained using cross-entropy loss with teacher forcing.
- Gradient accumulation is used to handle larger effective batch sizes.
- Training is performed for 10 epochs with a learning rate of 0.001.

3.4 Model 2 Transformer-based Seq2Seq

This is a Transformer-based Seq2Seq model, leveraging self-attention mechanisms for abstractive summarization.

3.4.1 Architecture

- **Encoder:** A multi-layer Transformer encoder processes the input sequence using self-attention. Each layer consists of multi-head attention and a position-wise feed-forward network.
- **Decoder:** A multi-layer Transformer decoder generates the output sequence using self-attention and cross-attention. It attends to both the input sequence and previously generated tokens.
- **Positional Encoding:** Added to input embeddings to provide positional information, as Transformers lack recurrence.
- **Multi-Head Attention:** Allows the model to focus on different parts of the input sequence simultaneously.
- **Output Layer:** A fully connected layer maps the decoder's output to the vocabulary size.

3.4.2 Implementation

- **Input Handling:** The input text is tokenized, converted into indices, and padded to the maximum sequence length.
- **Encoder:** The encoder processes the input sequence using self-attention and produces contextualized representations.
- **Decoder:** The decoder generates the output sequence using self-attention and cross-attention over the encoder's outputs.
- **Output Generation:** the model uses beam search to generate high-quality summaries during inference.
- **Training:** The model is trained using cross-entropy loss with the Noam optimizer and learning rate warmup.

3.4.3 Hyperparameters

Hyperparameter	**Value**
Batch Size	64
Embedding Dimension	128
Hidden Dimension	128
FFN Dimension	512
Number of Encoder Layers	3
Number of Decoder Layers	3
Number of Attention Heads	8
Dropout	0.3
Learning Rate	0.0001
Warmup Steps	4000

Table 2: Hyperparameters and their values

3.4.4 Training

- The model is trained using the Noam optimizer with a warmup schedule.
- Cross-entropy loss is used, and gradient clipping is applied to prevent exploding gradients.
- Training is performed for 15 epochs.

3.5 Model 3 Transformer-based Seq2Seq

3.5.1 Architecture, Implementation, Training

The architecture includes more dimensions, but the implementation and training are the same.

3.5.2 Hyperparameters

Hyperparameter	**Value**
Batch Size	32
Embedding Dimension	256
Hidden Dimension	256
FFN Dimension	1024
Number of Encoder Layers	3
Number of Decoder Layers	3
Number of Attention Heads	8
Dropout	0.1
Learning Rate	0.0003
Warmup Steps	4000

Table 3: Hyperparameters and their values

3.6 Model Differences

Aspect	Model 1	Model 2	Model 3
1. Architecture			
Encoder-Decoder	GRU-based	Transformer-based	Transformer-based
Attention Mechanism	Fixed attention	Multi-head self-attention	Multi-head self-attention
Positional Encoding	Not applicable	Yes	Yes
Complexity	Simpler, less computationally expensive	Moderate	High
2. Training			
Optimizer	Adam (fixed learning rate)	Noam optimizer (warmup schedule)	Noam optimizer (higher learning rate)
Teacher Forcing	Yes	Yes	Yes
Gradient Clipping	No	Yes	Yes
Gradient Accumulation	Yes	No	No
Training Stability	Moderate	High (due to self-attention and positional encoding)	Very high (larger capacity, optimized training)
3. Inference			
Beam Search	No	Yes	Yes
Summary Quality	Suboptimal	High-quality	Best quality (due to larger capacity)
4. Performance			
Sequence Handling	Struggles with longer sequences	Handles longer sequences better	Excels at complex patterns and longer sequences

Table 4: Differences Between Each Model

3.7 Plots

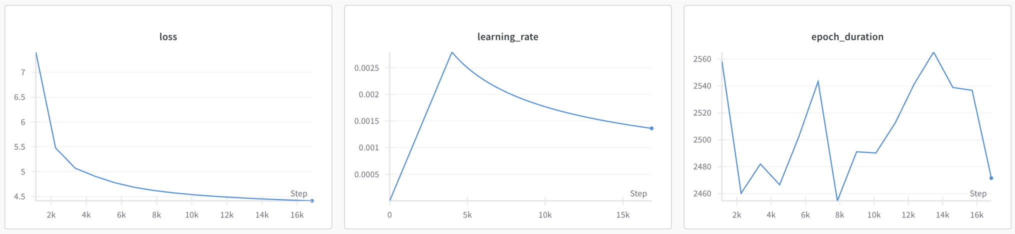

Figure 1: Model 2 Plots

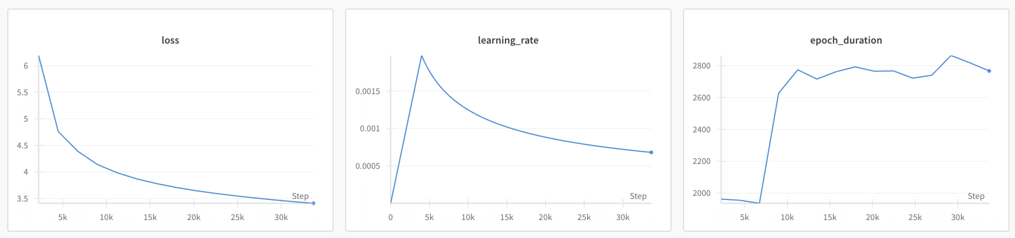

Figure 2: Model 3 Plots

3.8 Sample Generated Summaries

- **Model 2:**
  - The video shows the French authorities in the video of the plane crash. The video shows the plane crashed in the Air France. The plane crashed in the Air France, the plane crashed in the air. The plane crashed in the Air France, the plane crashed plane crashed in the air.
  - The UK has been eating food poisoning outbreak in the UK. The food poisoning outbreak has been linked to the deaths of the animals.
  - French President Francois Hollande has been sent to the French military base. The French military is believed to have been killed in the French military. The French military is believed to have been killed in the French military Mali.
  - `<unk>`, 46, was jailed for five years after she was jailed for five years.
  - Boy was found unconscious after he was hit by police after he

  was hit by police. Boy was found unconscious after he was pronounced dead. Police said he was 'bullied' and 'boy' boy was pronounced dead. Police said he was 'heartbroken' and ' ' 'heartbroken' and 'did not know the boy'.

  - Robert Durst, 30, was shot dead in the head in the head of his home in the Bronx. He was shot dead in the head and killed his wife, who was killed in the head. He was shot dead in the head and killed his wife, who was killed in the head and killed him. He was arrested on Tuesday after he was shot dead in the head and killed his wife.
  - The film was built in the attic. The film was inspired by the film.

- **Model 3:**
  - The United States is committed to the ruling. The United States is seeking a similar ruling in January. The United States and Israel signed a cease-fire warrant. The United States and Israel signed a cease-fire warrant Wednesday.
  - Amnesty International says international figures are ”not a threat.” China's international law enforcement official says the world is ”deeply disturbed.” The U.N. report says the number of executions is ”cruel and inhumane.” The U.N. report says the number of executions is at its highest in the world.
  - The Uyghur People's Association of Xinjiang Autonomous Region is seeking a new death penalty. The violence against the Uyghur people in Xinjiang region led to a death penalty. The government has been using the violence to attack the Uyghur people in China. The violence has sparked a widespread international outcry over the killing by the Chinese government.
  - ”Vampire Diaries” was announced Tuesday. Elena ¡unk¿ was ”Vampire.”
  - The number of people killed in the Mediterranean Sea is still unknown. The boat was headed to Italy on Saturday night. Authorities say the deaths are likely to be caused by the deaths of more than 700 migrants.

- Researchers at the University of London have developed a robot for 18 years. They are inspired by a variety of meals and are made up of them. They can then eat meals. They are then inspired by a variety of meals and even eat them.
- Louie the kitten is a growing trend in Washington D.C. The two-month-old kitten is seen in front of his family and adopted by his family. They are seen laughing as they look after their own cats and dogs.
- Nikki Kelly, 24, had been pregnant for three months. She had been pregnant for three months and had to be put together. She had to have a baby boy and had to be monitored.
- NASA's new ”smart” car is able to move on to space. The vehicle is at risk of electrical injuries, including electrical problems. NASA says it is ”very important” to move vehicles.

3.9 Results

The baseline GRU-based model gave really low scores, and also, the epochs took too long to run, thus adding the scores for the transformer-based models only.

Metric	Model 2	Model 3
ROUGE-1		
Recall (r)	0.1194	0.2335
Precision (p)	0.3261	0.3901
F1-Score (f)	0.1669	0.2838
ROUGE-2		
Recall (r)	0.0184	0.0527
Precision (p)	0.0431	0.0802
F1-Score (f)	0.0246	0.0612
ROUGE-L		
Recall (r)	0.1123	0.2140
Precision (p)	0.3081	0.3576
F1-Score (f)	0.1571	0.2602

Table 5: ROUGE Scores for Model 2 and Model 3

3.10 Conclusion

- **Model 1:**
  - Serves as a strong baseline with its GRU-based encoder-decoder architecture.
  - However, it struggles with longer sequences and capturing more complex patterns.
- **Model 2:**
  - Takes a big step forward by using a Transformer-based architecture.
  - Incorporates self-attention and beam search, which lead to better performance and more coherent summaries.
- **Model 3:**
  - Delivers the best results among the three models.
  - With larger embedding dimensions, optimized training, and increased capacity, it's perfect for high-quality summarization tasks.