**Model Architecture for Parent-Child Conversation Reconstruction**

**1. Overview**

The model aims to reconstruct fragmented online discussions, summarize them meaningfully, and evaluate their coherence. It consists of multiple components, each performing specific NLP tasks.

**2. System Components**

**2.1 Thread Reconstruction Module**

* **Graph-Based Approach**:
  + Uses networkx.DiGraph to model discussions as a directed graph.
  + Nodes represent comments (entry\_id), and edges represent parent-child relationships (parent\_id).
  + **Depth-First Search (DFS)** traverses the graph to reconstruct conversation threads.

**2.2 Summarization Module**

* **Model**: T5-base Transformer Model (pretrained from Hugging Face)
* **Input**: Concatenated reconstructed thread (limited to 512 tokens)
* **Output**: Compressed summary
* **Training/Fine-tuning**: Can be fine-tuned using cross-entropy loss.
* **Parameters**:
  + max\_length = min(150, 0.75 \* word\_count)
  + min\_length = min(5, 0.3 \* word\_count)
  + do\_sample = False (Ensuring deterministic summarization)

**2.3 Context Matching Module**

* **Model**: Sentence-BERT (all-MiniLM-L6-v2)
* **Function**: Compares comments to their parent comments to detect inconsistencies.
* **Similarity Metric**: Cosine similarity
* **Threshold**: Below 0.5 indicates a context mismatch

**2.4 Perplexity Estimation Module**

* **Model**: GPT-2 (gpt2 from Hugging Face)
* **Purpose**: Evaluates the fluency of the generated summaries
* **Loss Function**: Log-likelihood loss
* **Computation**:
  + Perplexity = exp(loss)
  + Avoids distorted calculations by handling NaN/infinite values.

**2.5 Performance Metrics**

1. **BLEU Score**: Measures n-gram overlap between original text and summary.
2. **ROUGE Score**: Evaluates key phrase recall in the summary.
3. **Perplexity**: Measures language model coherence.
4. **Semantic Similarity**: Assesses contextual retention in the summary.

**3. Training Objectives**

* **Reconstruct** meaningful threaded discussions from fragmented conversations.
* **Generate** coherent and concise summaries.
* **Detect** misleading or missing context using NLP techniques.
* **Evaluate** model outputs using standardized NLP metrics.

**4. Final Outputs**

The model produces:

* **Reconstructed conversations**
* **Generated summaries**
* **Mismatch detections**
* **Evaluation scores (BLEU, ROUGE, Perplexity, Similarity)**
* **Final processed dataset saved as CSV for further analysis.**

**Model Architecture for Parent-Child Conversation Reconstruction**

**1. Overview**

The model aims to reconstruct fragmented online discussions, summarize them meaningfully, and evaluate their coherence. It consists of multiple components, each performing specific NLP tasks.

**2. System Components**

**2.1 Thread Reconstruction Module**

* **Graph-Based Approach**:
  + Uses networkx.DiGraph to model discussions as a directed graph.
  + Nodes represent comments (entry\_id), and edges represent parent-child relationships (parent\_id).
  + **Depth-First Search (DFS)** traverses the graph to reconstruct conversation threads.

**2.2 Summarization Module**

* **Model**: T5-base Transformer Model (pretrained from Hugging Face)
* **Input**: Concatenated reconstructed thread (limited to 512 tokens)
* **Output**: Compressed summary
* **Training/Fine-tuning**: Can be fine-tuned using cross-entropy loss.
* **Parameters**:
  + max\_length = min(150, 0.75 \* word\_count)
  + min\_length = min(5, 0.3 \* word\_count)
  + do\_sample = False (Ensuring deterministic summarization)

**2.3 Context Matching Module**

* **Model**: Sentence-BERT (all-MiniLM-L6-v2)
* **Function**: Compares comments to their parent comments to detect inconsistencies.
* **Similarity Metric**: Cosine similarity
* **Threshold**: Below 0.5 indicates a context mismatch

**2.4 Perplexity Estimation Module**

* **Model**: GPT-2 (gpt2 from Hugging Face)
* **Purpose**: Evaluates the fluency of the generated summaries
* **Loss Function**: Log-likelihood loss
* **Computation**:
  + Perplexity = exp(loss)
  + Avoids distorted calculations by handling NaN/infinite values.

**2.5 Performance Metrics**

1. **BLEU Score**: Measures n-gram overlap between original text and summary.
2. **ROUGE Score**: Evaluates key phrase recall in the summary.
3. **Perplexity**: Measures language model coherence.
4. **Semantic Similarity**: Assesses contextual retention in the summary.

**3. Training Objectives**

* **Reconstruct** meaningful threaded discussions from fragmented conversations.
* **Generate** coherent and concise summaries.
* **Detect** misleading or missing context using NLP techniques.
* **Evaluate** model outputs using standardized NLP metrics.

**4. Final Outputs**

The model produces:

* **Reconstructed conversations**
* **Generated summaries**
* **Mismatch detections**
* **Evaluation scores (BLEU, ROUGE, Perplexity, Similarity)**
* **Final processed dataset saved as CSV for further analysis.**