Chunking: All About Boutique Models

In boutique models, which serve as alternatives to web search and traditional LLMs, chunking is not merely a preprocessing step but a pivotal process that determines the overall model performance and cost structure.

Dec 7 2024 / Sangzun Park

1. From Words Count to Semantic Chunking

Characteristics of Words Counts Chunking

Words Counts Chunking involves dividing text based on the number of words, creating chunks of a specific word count.

- **Example:**
  If the text is "The cat sat on the mat", and you chunk it by two words, it becomes ["The cat", "sat on", "the mat"].
- **Features:**
  - **Simple and Efficient:** Easy to implement and consumes minimal computational resources.
  - **Lack of Context Awareness:** Since it chunks text purely based on word count, it fails to capture contextual meaning.

Characteristics of Semantic Chunking

Semantic Chunking takes into account the semantic relationships and context of words when chunking text. This approach allows for deeper processing of information by segmenting text based on semantic connections.

- **Example:**
  For the text "The cat sat on the mat to rest", semantic chunking may divide it as ["The cat sat on the mat", "to rest"], reflecting the context.
- **Features:**
  - **Context-Aware:** Considers the semantic structure of the text rather than simple word counts.
  - **Complex Processing:** Requires more advanced algorithms and higher computational resources.

Advantages of Shifting from Words Counts Chunking to Semantic Chunking

1. **Enhanced Contextual Meaning**
   Semantic Chunking goes beyond simple word sequences to reflect semantic relationships within sentences, enabling more in-depth analysis and natural outputs.
2. **Improved Text Processing Accuracy**
   By structuring data based on meaning, Semantic Chunking improves accuracy in applications like text summarization and information retrieval.
3. **Semantic Integration**
   Focuses on meaning rather than word order or placement, allowing for uniform processing of varied expressions with the same meaning.
    - **Example:** Treating "buying a car" and "purchasing an automobile" as the same chunk.

Disadvantages of Shifting from Words Counts Chunking to Semantic Chunking

1. **Increased Computational Costs**
   Semantic Chunking increases the complexity of natural language processing models, requiring more computational resources and time.
2. **Dependency on Accuracy**
   The performance of Semantic Chunking heavily depends on the quality of the underlying model or algorithm. Poorly trained models can produce inaccurate results.
3. **Implementation Complexity**
   Semantic Chunking is harder to implement and maintain compared to Words Counts Chunking.

2. Some Methods for Semantic Chunking

1. SentenceTransformer

- **Description:**
  A BERT-based embedding model that vectorizes sentences to calculate similarity or understand context. It performs chunking by measuring semantic similarity between sentences, often using cosine similarity to define chunk boundaries.

Advantages

- **Context Awareness:**
  Effectively identifies semantic relationships between sentences, making it suitable for grouping similar sentences into the same chunk.
- **High Processing Speed:**
  Pre-trained models enable fast computation and easy batch processing.
- **Diverse Pre-Trained Models:**
  Offers flexibility with models fine-tuned for various domains.

Disadvantages

- **Single-Sentence Limitation:**
  Processes semantic relationships between sentences relatively simply and struggles with complex context understanding.
- **Fixed Pre-Trained Model Constraints:**
  Requires additional fine-tuning for domain-specific chunking tasks.
- **Limits in Semantic Interpretation:**
  May fail to grasp subtle nuances or metaphorical expressions accurately.

2. Chonkie

- **Description:**
  An open-source chunking library that uses predefined rules and patterns to chunk text. It operates by separating data based on specific keywords, grammar, or sentence structure.

Advantages

- **Customizable:**
  Rule-based operation allows for customization according to specific text structures or patterns.
- **Strength in Structured Data:**

Shows high accuracy when data follows a consistent pattern.

- **Lightweight:**
  Operates quickly and efficiently without requiring external API calls.

Disadvantages

- **Lack of Flexibility:**
  Struggles to handle exceptional cases due to its rule-based nature.
- **No Contextual Understanding:**
  Relies solely on form and patterns, ignoring semantic relationships.
- **Complex Rule Design:**
  Designing rules to cover diverse scenarios can be challenging and difficult to maintain.

3. LLMs (GPT, Mistral, etc.)

- **Description:**
  Large language models perform chunking by deeply understanding the context and semantics of text. Using prompts, users can directly guide the chunking process. Chunking methods can involve simply leveraging the LLM for semantic chunking or creating a structured process to execute chunking.

Advantages

- **High Context Understanding:**
  Effectively comprehends complex context and semantics for flexible chunking.
- **Domain Adaptability:**
  Can adapt to specific domains or styles by adjusting or directing chunking tasks.
- **Creative Processing:**
  Easily integrates with additional tasks like text generation or summarization beyond basic chunking.

Disadvantages

- **Cost:**
  API calls incur costs, making it economically challenging for large-scale data processing.
- **Processing Speed:**
  Slower than SentenceTransformer or Chonkie due to the large-scale model architecture.
- **Unstable Results:**
  May produce inconsistent results for the same input, requiring additional validation for accuracy.

Comparison Summary

Method	Description	Advantages	Disadvantages
SentenceTransformer	Embeds sentences and performs similarity-based chunking	- Fast and efficient - Offers diverse domain models - Captures contextual similarity	- Limited understanding of complex context - Domain-specific tuning needed - Struggles with metaphors and nuances
Chonkie	Rule-based chunking	- Easy customization - Suitable for structured data - Lightweight and fast	- Lacks flexibility - Ignores semantic relationships - Rule design and maintenance can be challenging
LLM / Customized LLM	Contextual chunking using large language models	- High contextual understanding - Adaptable to specific domains - Integrates with additional tasks	- High costs - Slow processing speed - Inconsistent results

Results

1. Limitations and Advantages of Rule-Based Chunking

- Rule-based libraries like Chonkie consume fewer resources during execution and are suitable for specific text structures.
- Require significant resources during the preparation phase and need constant adjustments depending on the text structure and nature.

2. Strengths and Weaknesses of SentenceTransformer

- Easily provides above-average results and includes models tailored for specific domains.
- Limited in understanding complex contexts.

3. LLM-Based Chunking Methods

- Relying on the basic capabilities of LLMs for simple chunking.

- Customizing chunking with prompts and additional coding.
- The basic method is easy to implement but yields inconsistent results.
- Customization provides stability and higher accuracy but requires significant time and resources.

4. Key Implementation and Challenges

- **Context Merging:**
  - Focused on incrementally merging chunks with similar contexts.
  - Faced issues with oversized chunks that were not manageable with current LLM APIs.
- **Attempts to Resolve:**
  - Tried various methods like chunk size limitations and batch processing, but these consumed substantial time and revealed inefficient structures.
  - Text loss also became an issue.

5. Optimized Solution

- **Chunk Splitting:**
  - Adopted a method to split oversized chunks into smaller ones, overcoming LLM token size limitations and preventing text loss.
- **Outcome:**
  - Using Mistral AI, processing a 50-page PDF took over 5 hours but produced the most satisfactory results so far.

6. Future Directions

- **Further Validation:**
  - Plan to validate the efficiency and accuracy of the current method.
- **Designing a Hybrid Architecture:**
  - Aim to explore hybrid architectures that combine rule-based chunking libraries with LLMs to address various text processing challenges effectively.

3. Evaluation Methods

Evaluation Method	Description	Advantages	Disadvantages
Cosine Similarity-Based Evaluation	Calculates similarity using text embedding vectors	- Simple to implement - Fast computation - Handles linguistic variations	- Limited reflection of context and semantic nuances - Relies on embedding models - Suitable for approximate similarity
ROUGE-Based Evaluation	Scores based on word overlap and alignment information	- Fast and straightforward - Enables quantitative comparison - Widely used	- Lacks context awareness - Weak at detecting hallucinations - Limited effectiveness with long texts
Sliding Window + NLI	Compares context and evaluates entailment on a window basis	- Verifies contextual alignment - Allows detailed validation - Handles large datasets	- High computational resource consumption - Dependence on models - Limitations in processing long sentences
LLM-Based Evaluation	Performs semantic evaluation of sentences using GPT	- High contextual understanding - Provides detailed explanations - Adapts well to specific domains	- Expensive - Slower processing - Inconsistent results

Results

1. Need for a Hybrid Approach

- Cosine similarity offers consistency but often fails to identify the most contextually appropriate answer, as it prioritizes minimal errors over meaning.
- LLM-based evaluation excels at capturing subtle nuances and context but lacks the stability of cosine similarity.

2. **Importance of Human Evaluation**
    - Regardless of the method, human evaluation is essential to ensure accurate results.
    - The quality of human evaluation depends on the evaluator's language proficiency, logical reasoning, and domain expertise.
3. **Conclusion**
    - A balanced combination of cosine similarity and LLM-based evaluation, complemented by human oversight, is the most effective approach for reliable assessment.