DISTILL: Detailed Intelligent Summarization for Text Information Learning and Linkage

Author - Suphakit Ng
Advisor - Associate Professor Dr.Jakramate Bootkrajang

ABSTRACT

This research presents DISTILL, a framework combining semantic analysis with graph-based knowledge representation to enhance document summarization. Current summarization approaches often lose critical relationships between concepts and fail to maintain the hierarchical structure of academic documents, potentially misrepresenting key findings and their supporting evidence. DISTILL addresses these challenges through a multi-stage pipeline of semantic chunking, knowledge graph construction, and summary generation, preserving both semantic relationships and document structure in the final output.

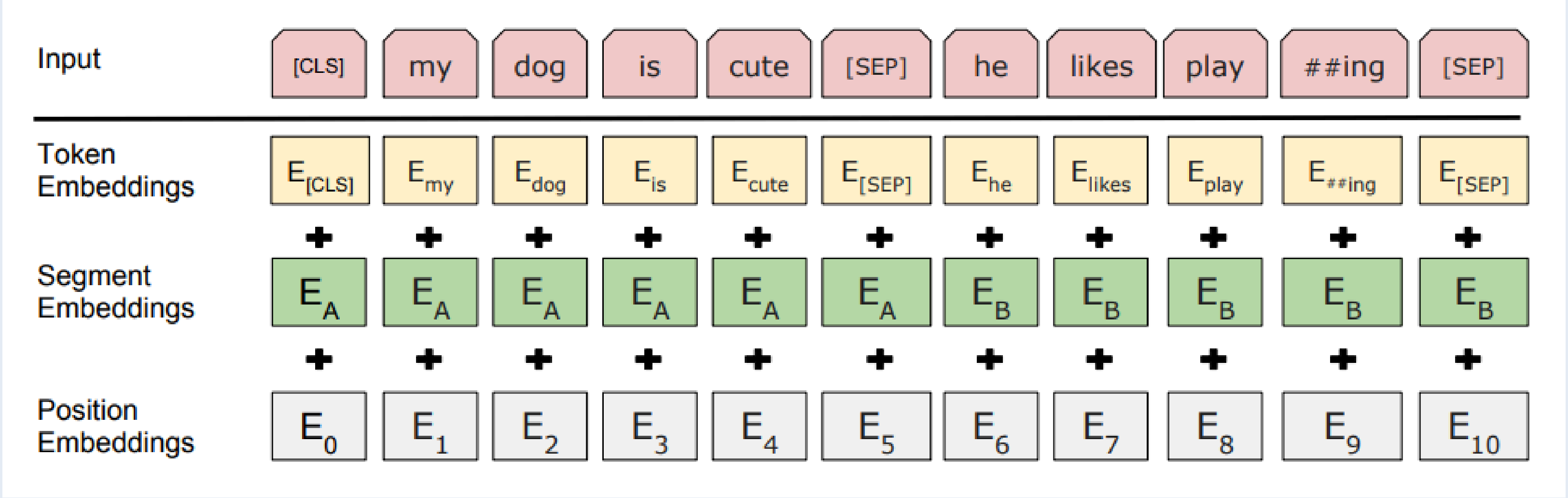
INTRODUCTION

Document summarization has become increasingly critical in managing the growing volume of academic literature, yet current approaches often fail to capture the intricate web of relationships between concepts, claims, and evidence that form the backbone of academic discourse. DISTILL addresses these limitations through a graph-based approach that weights and preserves semantic relationships. By integrating BERT embeddings with graph-based knowledge representation, DISTILL creates a semantic framework that understands both local relationships between concepts and global document structure, enabling summaries that maintain the intellectual integrity of academic works while preserving their argumentative structure.

METHODOLOGY

Semantic Text Processing

For semantic text processing, DISTILL adopts the BERT embedding approach similar to Devlin et al. (2019), utilizing its contextual embeddings for semantic chunking. However, unlike traditional BERT applications, DISTILL implements a modified semantic similarity threshold mechanism (0.95) to maintain coherent document segments while preserving cross-sectional relationships.


Input
[CLS]
my
dog
is
cute
[SEP]
he
likes
play
##ing
[SEP]
Token Embeddings
$E_{[CLS]}$ E_{my} E_{dog} E_{is} E_{cute} $E_{[SEP]}$ E_{he} E_{likes} E_{play} $E_{\#\#ing}$ $E_{[SEP]}$
Segment Embeddings
E_A E_A E_A E_A E_A E_A E_B E_B E_B E_B E_B
Position Embeddings
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}

Ref: Devlin et al., 2019

Graph Construction

The graph construction phase draws inspiration from TextRank (Mihalcea and Tarau, 2004), but extends beyond basic keyword extraction. DISTILL incorporates a hierarchical graph structure similar to Liu & Lapata's (2019) approach, while introducing a concept weighting system. Each concept's importance is calculated using a weighted formula shown below.

$$I_c = 0.3F + 0.2K + 0.3S + 0.2T$$

I_c is the importance score of concept c

$F = \frac{f_c}{N}$ is the normalized frequency of concept c in document

$K = \{ \begin{matrix} 1 \text{ if concept contains NOUN/PROPN} \\ 0 \text{ otherwise} \end{matrix}$

$S = \{ \begin{matrix} 1 \text{ if concept is subject} \\ 0 \text{ otherwise} \end{matrix}$

$T = \{ \begin{matrix} 1 \text{ if concept has technical term} \\ 0 \text{ otherwise} \end{matrix}$

The system also employs SpaCy's NLP library for entity recognition, which enables accurate identification of concepts and their syntactic roles for weight calculation

Summarization

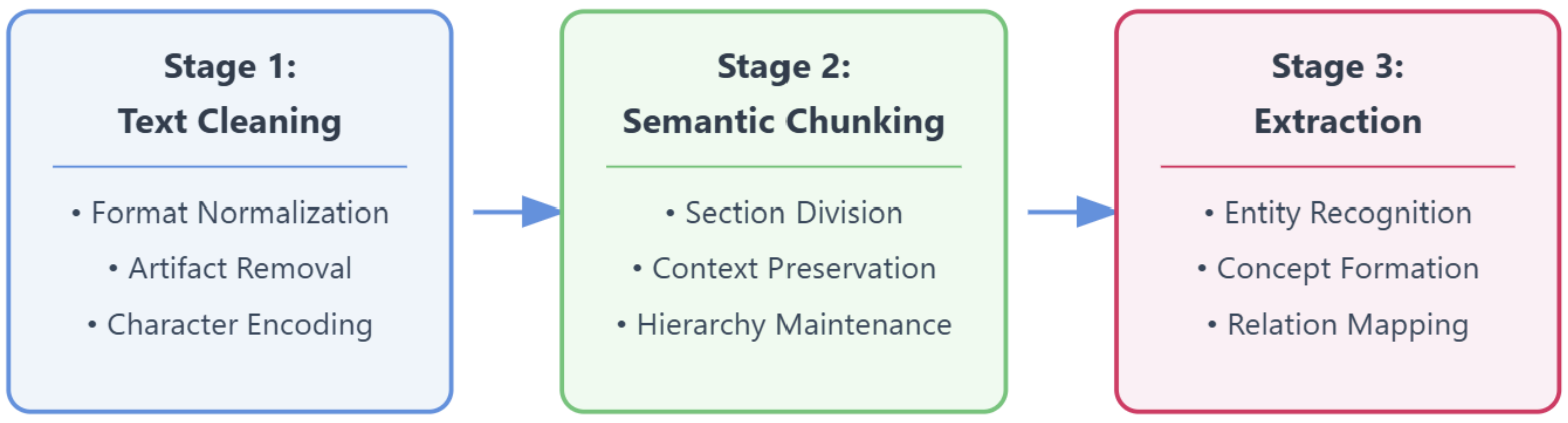
For summarization, DISTILL uses the Falconsai/text_summarization model as a baseline traditional summarizer. This pretrained model provides a benchmark for comparing the effectiveness of DISTILL's graph-based approach.

DATA - PREPROCESSING

DISTILL is primarily designed to process and analyze research papers in PDF format, focusing on extracting structured information from complex academic documents. DISTILL implements a three-stage preprocessing pipeline to transform raw documents into structured, analyzable content:

1. Text Cleaning transforms raw documents by normalizing formats and removing conversion artifacts, ensuring clean and consistent text output.
2. Semantic Chunking divides documents into coherent sections while maintaining contextual relationships between segments.
3. Extraction Phase identifies and organizes key textual elements, preparing structured content for knowledge graph construction.

Throughout this pipeline, DISTILL prioritizes document structure preservation, ensuring the integrity of section relationships and contextual information.



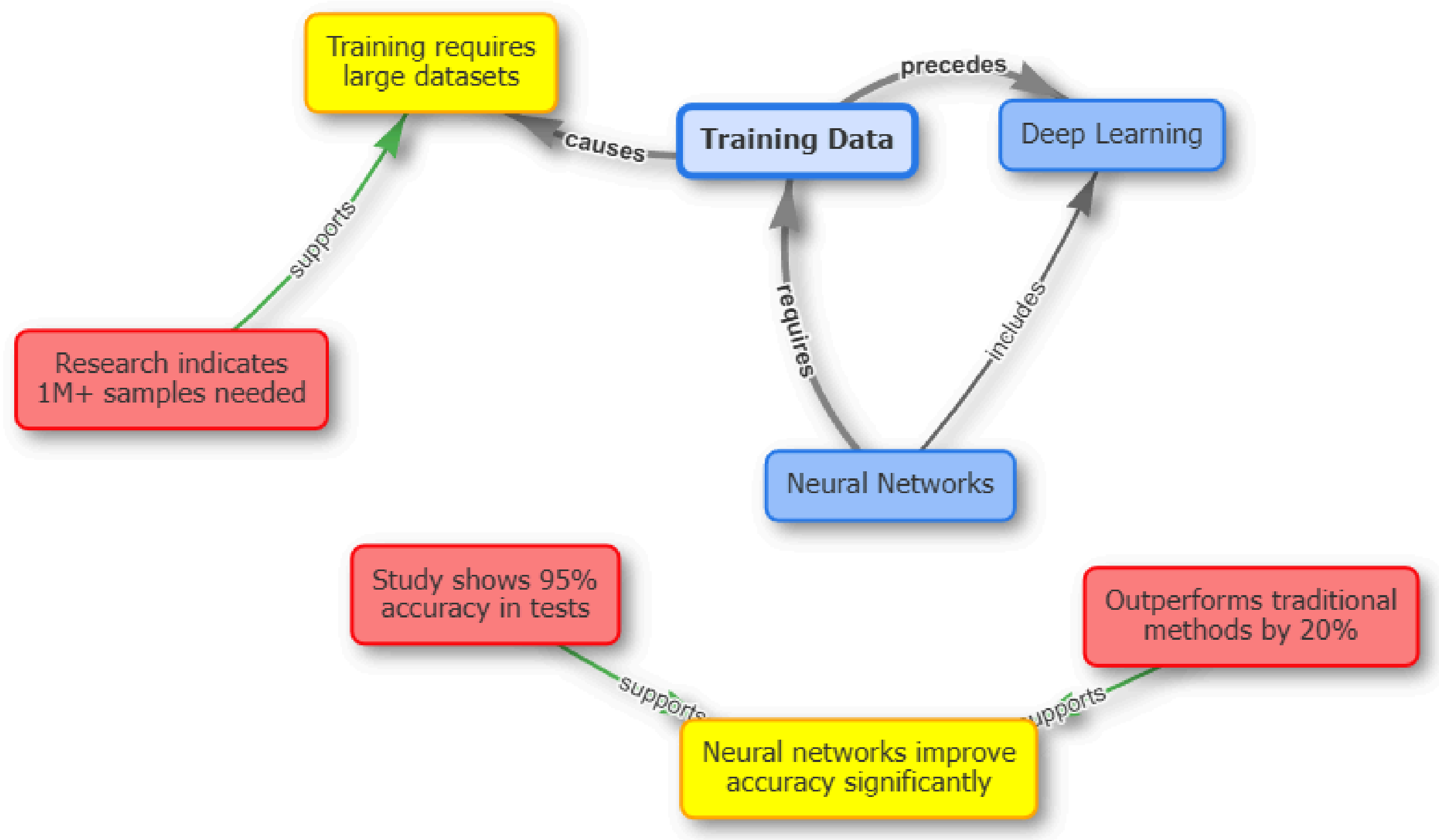
Stage 1:
Text Cleaning
- Format Normalization
- Artifact Removal
- Character Encoding

Stage 2:
Semantic Chunking
- Section Division
- Context Preservation
- Hierarchy Maintenance

Stage 3:
Extraction
- Entity Recognition
- Concept Formation
- Relation Mapping

Exploratory Data Analysis

After semantic chunking and BERT embedding generation, we need to visualize document structure. Using graph-based visualization techniques, the system plots concept relationships and claim-evidence chains, enabling clear representation of semantic connections between document sections as shown in the figure below.


Training requires large datasets
causes
Training Data
precedes
Deep Learning
supports
Research indicates 1M+ samples needed
requires
includes
Neural Networks
Study shows 95% accuracy in tests
Outperforms traditional methods by 20%
supports
supports
Neural networks improve accuracy significantly

Evaluation

The evaluation framework integrates standard metrics (ROUGE, BERTScore), drawing from Zhang et al.'s (2020). This approach enables comprehensive evaluation of both summary quality and semantic retention.

Results & Discussion

The evaluation was conducted using 100 samples from the CNN/DailyMail dataset. The results demonstrate DISTILL's superior performance across multiple metrics. While both approaches show competitive ROUGE scores, DISTILL achieves notably higher results with ROUGE-1 (**0.51** vs 0.42), ROUGE-2 (**0.40** vs 0.39), and ROUGE-L (**0.45** vs 0.41). Although the baseline model shows slightly higher cosine similarity (0.80 vs 0.74), DISTILL's graph-based approach generates more semantic match summaries compared to direct summarization model.

Metric	Traditional	Graph-based
ROUGE-1	0.422 ± 0.041	**0.511 ± 0.088**
ROUGE-2	0.389 ± 0.044	**0.405 ± 0.096**
ROUGE-L	0.406 ± 0.045	**0.448 ± 0.088**
BERTScore	**0.276 ± 0.086**	0.094 ± 0.113
Cosine Similarity	**0.803 ± 0.074**	0.743 ± 0.116