


# Enhancing Multi-Document Summarization with Cross-Document Graph-based Information Extraction

**Presented by:**  
**Kushan Hewapathirana – 229333P**

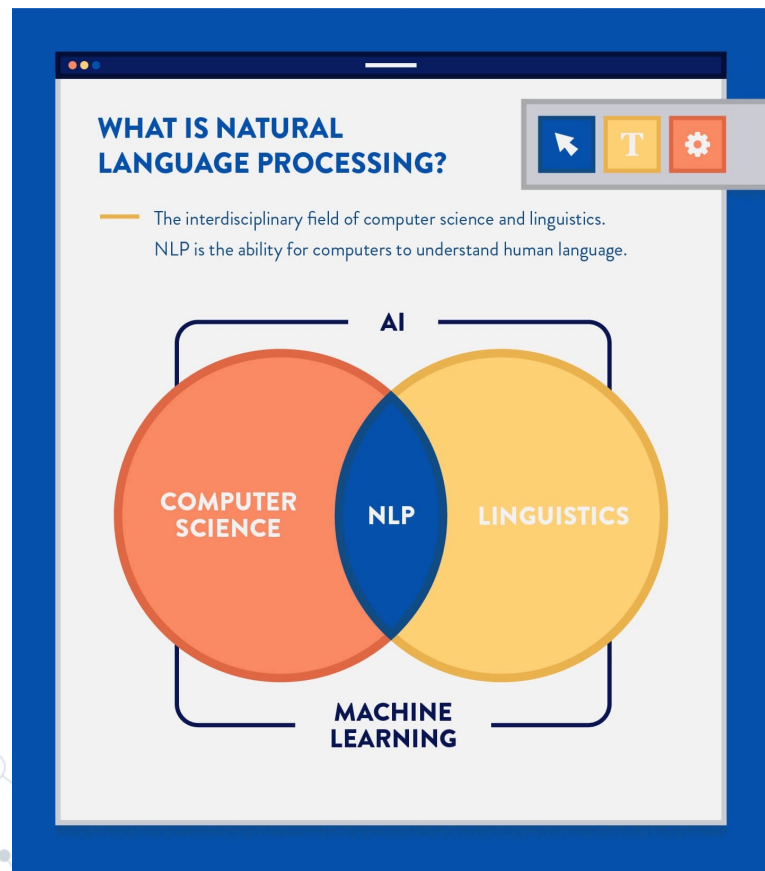
**Published in:** EACL 2023: Main Volume: May 2023





# **APPLICATION DOMAIN: Natural Language Processing – Multi-document Summarization**

# Introduction to Natural Language Processing



**Speech  
recognition**

**Part of speech  
tagging**

**Word sense  
disambiguation**


**Named entity  
recognition**

**Co-reference  
resolution**

**Sentiment  
analysis**

**Natural  
language  
generation**

# Introduction to Information Extraction (IE)



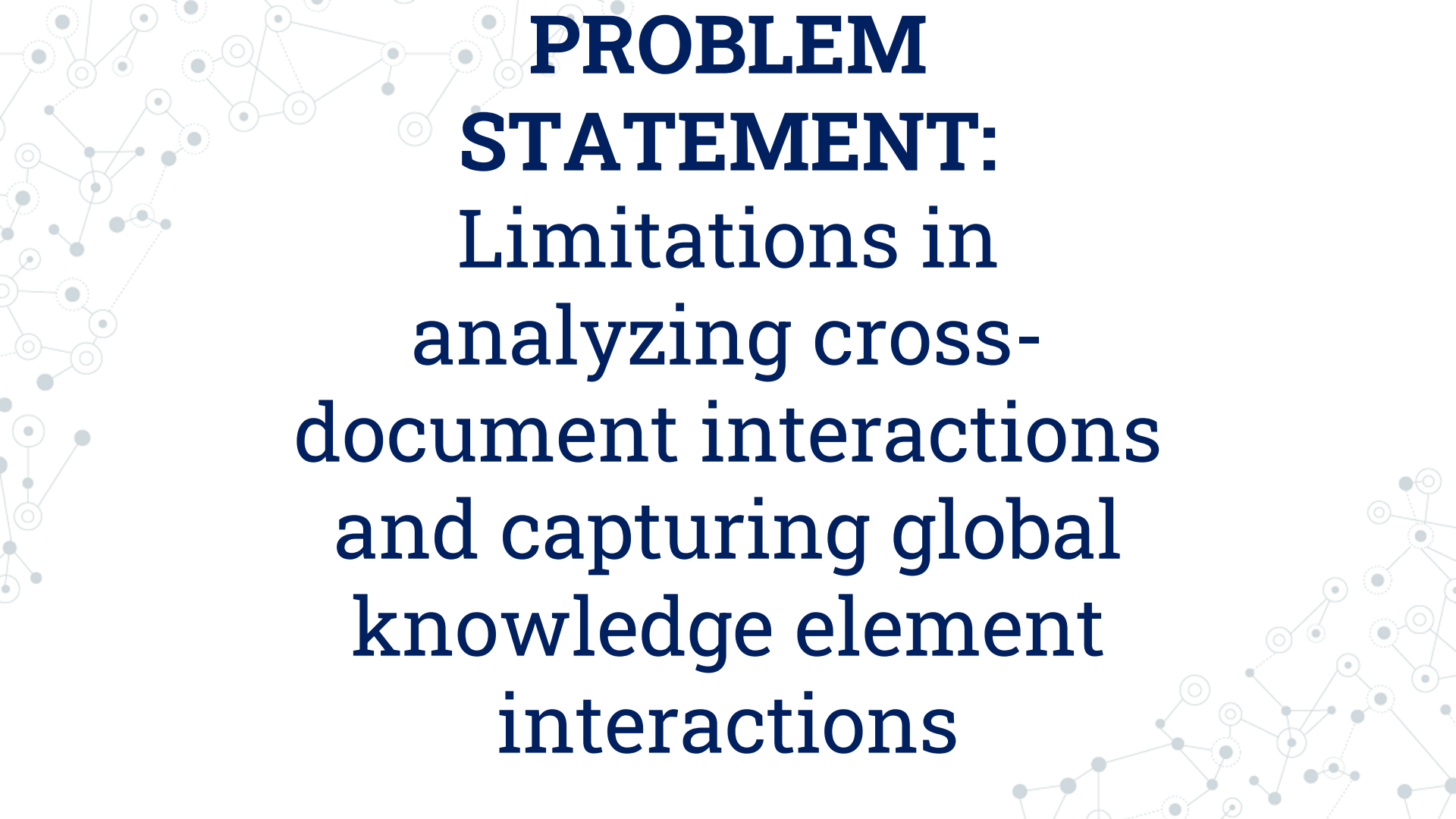
Process of automatically extracting structured information from unstructured or semi-structured documents

Identifying and extracting specific pieces of information, such as entities, relationships, and attributes

Applied to a wide range of textual sources, including emails, web pages, reports, presentations, legal documents, and scientific papers

# Techniques and Methods used for Information Extraction

- **Named Entity Recognition (NER)**: Identifying and classifying named entities in text into pre-defined categories such as person names, organization names, and location names.
- **Relationship Extraction** : Identifying and extracting relationships between entities in a text.
- **Deep learning**: Advanced deep learning architectures, including recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer models like BERT.
- **Dependency Parsing** : Analyze the grammatical structure of a sentence and extract relationships between words, which can be helpful in identifying the relationships between entities.
- **Text Summarization** : Creating a summary of a text by either extracting parts of the text or generating fresh text that conveys the crux of the original text



# **PROBLEM STATEMENT:** Limitations in analyzing cross- document interactions and capturing global knowledge element interactions

# Challenges in Information Extraction



Inconsistencies between IE graphs and text representations.

Cross-document Information Extraction.

Capturing cross-document interactions and analysing global interactions.

# Bridging the Research Gap: Unique Contribution

Inability to capture cross-document interactions and analyse global interactions among extracted knowledge elements

- The paper proposes a cross-document fine-grained IE system to extract a cluster-level information graph.

IE is focused on extracting structured information, whereas MDS aims to condense the most important information into a natural language summary. There has been limited research using IE techniques to improve MDS

- This paper proposes a text summarization model enhanced by IE that focuses on MDS and improves the MDS model using cross-document IE graphs.
- Construct the entire graph to capture global interactions among the extracted knowledge elements.

Hallucination, a technical limitation of text generation methods, which affect the accuracy and reliability of extracted information

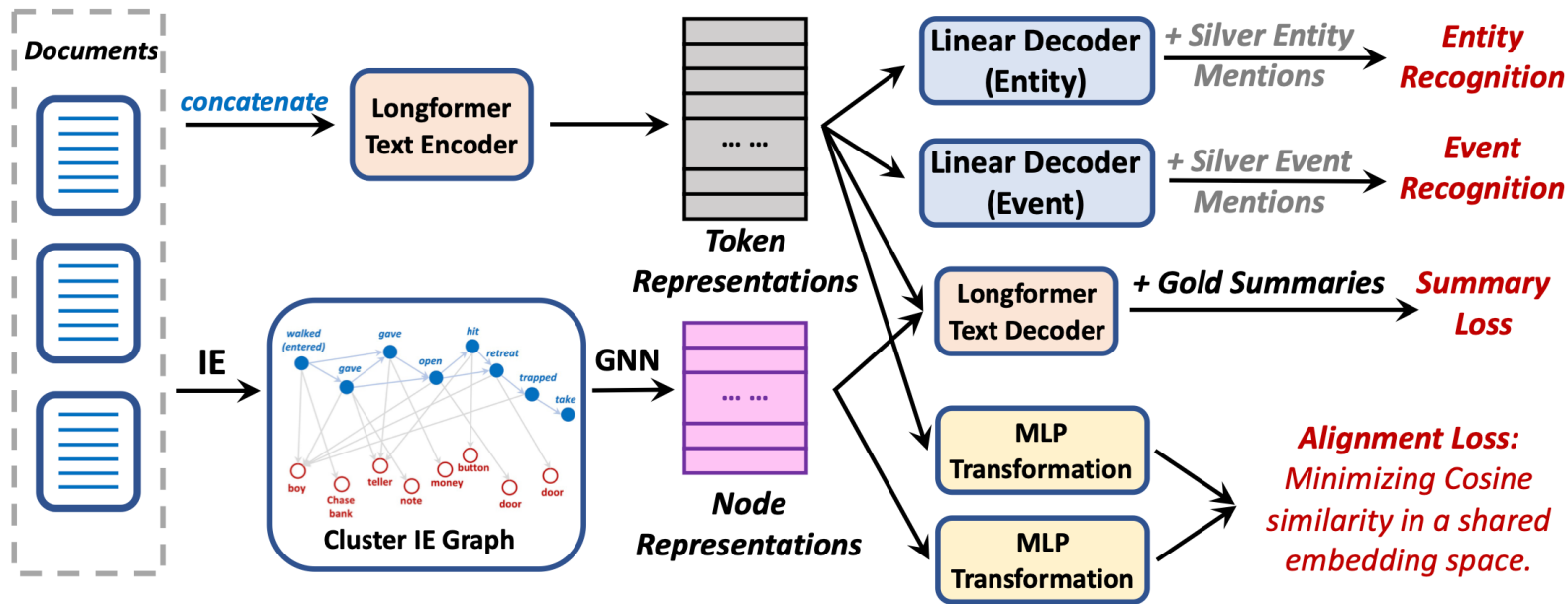
- The proposed approach improves the factual consistency of generated summaries with the source documents while maintaining the same level of abstractiveness.





# **Methodology:** Enhancing MDS by using cross-document graph-based IE

# Proposed Approach

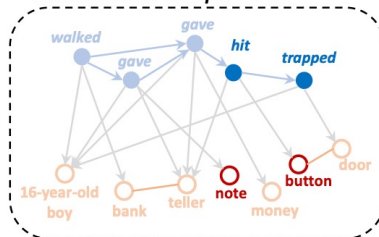


# An Example Of An Extracted Cross-document IE Graph

## Document #1:

... Detroit Police Officer Dan Donakowski said the 16-year-old boy walked into the bank located in the 15000 block of West 7 Mile around 2 p.m. and demanded money. Donakowski said the teen gave the teller a note and threatened to use a bomb if she didn't fork over the cash. "The teller complied, gave him some money and as he attempted to leave, the teller hit the button for the doors that automatically lock and the suspect was trapped inside," Donakowski said.

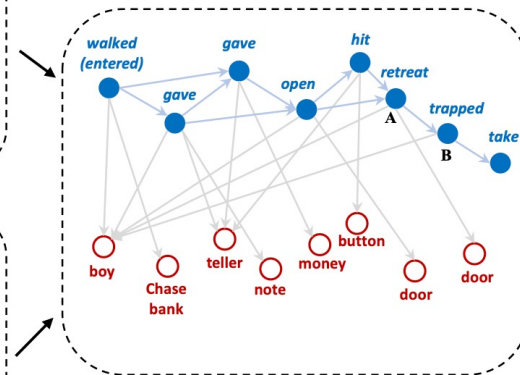
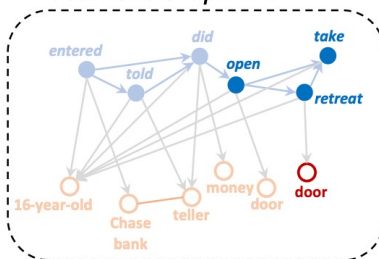
## Document Graph #1:



## Document #2:

... Police say a 16-year-old, of Detroit, entered Chase Bank located on Seven Mile on Detroit's east side about 2:30 p.m. Monday. He walked up to the counter and told the teller he was strapped with a bomb and to give him all the money. The teller did. The teen set off for the doorway. He opened the first set of doors into the causeway. The sidewalk was only steps away. He made it to the outermost set of doors, inches from the outside world. He'd make it no further. After realizing they wouldn't budge, he tried to retreat through the door he'd just passed. They wouldn't budge either. police can arrive and take him safely into custody ...

## Document Graph #2:



IE graph for the document cluster

# Cross-Document Information Extraction

- Cross-document information extraction is performed on each document cluster using two state-of-the-art systems: ReFinED [1] for entity extraction and disambiguation, and RESIN-11 [2] for event extraction and tracking.
- Relation extraction, event argument role labeling, and event-event temporal relation extraction are performed to add edges and obtain a complete IE graph for each document
- To connect document-level IE results into a cross-document IE graph, cross-document entity and event coreference resolution is performed

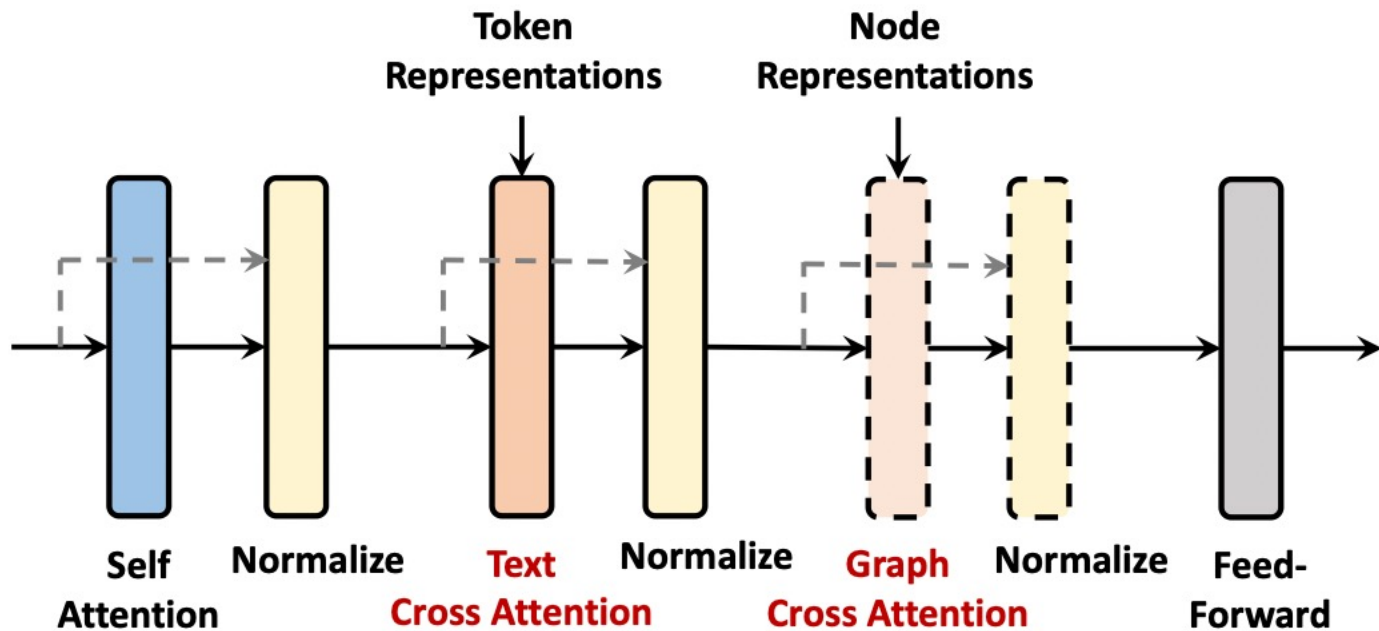
[1] Tom Ayoola, Joseph Fisher, and Andrea Pierleoni. 2022. Improving entity disambiguation by reasoning over a knowledge base. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2899–2912, Seattle, United States. Association for Computational Linguistics.

[2] Xinya Du, Zixuan Zhang, Sha Li, and others. 2022. RESIN-11: Schema-guided event prediction for 11 newsworthy scenarios. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations, pages 54–63, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

## Document Encoding

- The document encoding in the PRIMERA model is done using the encoder of the pre-trained Longformer-Encoder-Decoder (LED) model.
- Global attention is assigned to the document boundary tokens to analyze the relationships between the documents.
- Cross-document IE system is used to extract a cross-document IE graph
- The entity and event nodes in the graph are encoded using an edge-conditioned graph attention network.

# Summary Generation - Decoder Layer



# Recognizing Entities and Events, and Aligning Nodes and Text

- The model uses entity and event recognition to improve summarization by identifying important events and entities.
- A Multi-Layer Perceptron (MLP) classifier is used to classify tokens into entity, event, or none categories.
- The entity and event recognition loss is computed based on the correct labels for each token.
- Node and text alignment is proposed to ensure coordination between the graphs and summarization text.
- The alignment loss minimizes the distance between node representations and their corresponding texts to reduce errors and noise from the extraction system.

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by small circles, some of which are larger and have concentric rings, suggesting different levels of connectivity or importance. The lines are thin and gray, creating a mesh-like structure.

# **EXPERIMENTS AND RESULTS**

A decorative network diagram in the bottom-right corner, similar to the one in the top-left. It shows a cluster of nodes connected by lines, with some nodes being larger and more prominent than others. The overall style is clean and modern, using a light gray color scheme.



# Experimental Setup



## Data:

Datasets used: news articles from Multi-news, Wikipedia articles from WCEP-10, and DUC-2004.



## Baselines:

various unsupervised extractive and abstractive summarization models, including PRIMERA, REFLECT, BART-Graph, and LED.



## Experimental Process:

The researchers enhanced LED model using IE and employed simple summarization baselines, including PRIMERA, PRIMERA, REFLECT, BART-Graph, and LED.



## Experimental Environment:

on 8 NVIDIA V100 GPUs with 32GB memory.



## Statistics of the MDS Datasets

Dataset	# Train / Val / Test	Docs per Cluster	Average Summary Length
Multi-News	44972 / 5622 / 5622	2.8	217
WCEP-10	8158 / 1020 / 1022	9.1	28
DUC-2004	0 / 0 / 50	10	115

# Test Results

Evaluation Metrics	Co-occurrence			Factual Consistency					Abtractiveness
	R-1	R-2	R-L	FactCC	FactGraph	SUMMAC	Bert-P	EntityPrec	MINT
<i>Multi-News</i>									
<i>REFLECT</i>	49.3	20.0	24.8	-	-	-	-	-	-
<i>BART-Graph</i>	49.2	19.0	24.0	74.2	74.1	86.0	<b>87.3</b>	89.9	81.8
<i>PRIMERA</i>	49.9	20.9	25.8	73.1	75.0	86.2	87.0	89.3	82.1
Separate-Graphs	49.8	20.4	25.8	74.7	75.2	86.2	87.0	89.5	82.1
Recognition-Only	50.0	20.8	26.0	77.4	76.1	86.5	87.1	<b>91.1</b>	82.0
Alignment-Only	50.3	20.9	26.3	75.6	74.9	87.7	87.1	90.8	82.1
<b>Full Model</b>	<b>50.3</b>	<b>21.1</b>	<b>26.4</b>	<b>77.8</b>	<b>76.5</b>	<b>87.9</b>	87.1	<b>91.1</b>	82.1
<i>WCEP-10</i>									
<i>PRIMERA</i>	46.1	24.9	37.8	68.0	71.3	56.9	94.1	88.0	86.6
Separate-Graphs	46.1	24.8	37.8	69.1	71.6	57.0	94.0	89.1	86.6
Recognition-Only	46.1	24.8	<b>37.9</b>	71.2	<b>71.7</b>	57.1	94.0	91.0	86.5
Alignment-Only	<b>47.3</b>	<b>25.0</b>	<b>37.9</b>	68.5	71.4	57.6	<b>94.4</b>	90.5	86.5
<b>Full Model</b>	<b>47.3</b>	24.9	37.8	<b>71.5</b>	<b>71.7</b>	<b>57.7</b>	<b>94.4</b>	<b>91.3</b>	86.8
<i>DUC-2004</i>									
<i>PRIMERA</i>	32.6	6.7	16.8	53.0	48.8	77.9	84.2	79.6	70.1
Separate-Graphs	32.6	6.6	16.8	54.2	49.9	77.6	<b>85.1</b>	80.4	70.1
Recognition-Only	32.5	6.8	16.8	54.2	<b>51.2</b>	76.8	84.7	82.3	70.4
Alignment-Only	32.8	<b>7.2</b>	17.1	53.2	49.1	78.9	84.3	80.0	70.2
<b>Full Model</b>	<b>32.9</b>	<b>7.2</b>	<b>17.3</b>	<b>54.8</b>	<b>51.2</b>	<b>79.1</b>	85.0	<b>84.0</b>	70.1

## Test Results - Human Evaluation Results

Dataset	Major (1.0)	Minor (2.0)	No (3.0)	Avg Scores
baseline	6.0%	11.3%	82.7%	2.767
Multi-News (ours)	<b>4.7%</b>	12.7%	82.7%	<b>2.780</b>
baseline	10.7%	9.0%	80.3%	<b>2.697</b>
WCEP-10 (ours)	<b>9.0%</b>	19.3%	71.7%	2.627
baseline	22.0%	22.0%	56.0%	2.340
DUC-2004 (ours)	<b>18.0%</b>	12.0%	70.0%	<b>2.520</b>

# Test Results - A Comparison of the PRIMERA and Proposed Method

## *Input Documents:*

SAN FRANCISCO (MarketWatch) -- Trading in all Nasdaq-listed stocks and options was halted on Thursday due to technical problems on the bourse, according to Nasdaq OMX Group (NASDAQ:NDAQ).... In response, the New York Stock Exchange has also stopped trading in all Nasdaq securities at the request of Nasdaq OMX. "All orders in those securities have been canceled back to customers," said NYSE in a statement. **The Nasdaq Composite index (NASDAQ:COMP) was last at 3631.17, up 31.38 points**, before trading was suspended. There was no immediate word on when transactions will resume...

Article Excerpt A technical glitch knocked out trading in all Nasdaq Stock Market securities for **three hours Thursday** afternoon, an unprecedented meltdown for a U.S. exchange that paralyzed a broad swath of markets and highlighted the fragility of the financial world's electronic backbone. Nasdaq officials scrambled to figure out what happened and resume trading. They shared few of their findings with trading firms or the public during regular trading hours, ...

## *Summary (without Recognition & Alignment)*

Trading in all Nasdaq-listed stocks and options was halted **on Thursday** due to technical problems on the bourse, MarketWatch reports. The exchange sent out a series of emails alerting investors that it was experiencing issues with "quote submissions." In response, the New York Stock Exchange has also stopped trading in all Nasdaq securities at the request of Nasdaq OMX. "All orders in those securities have been canceled back to customers," the **exchange** said in a statement. There was no immediate word on when transactions will resume.

## *Summary (with Recognition & Alignment)*

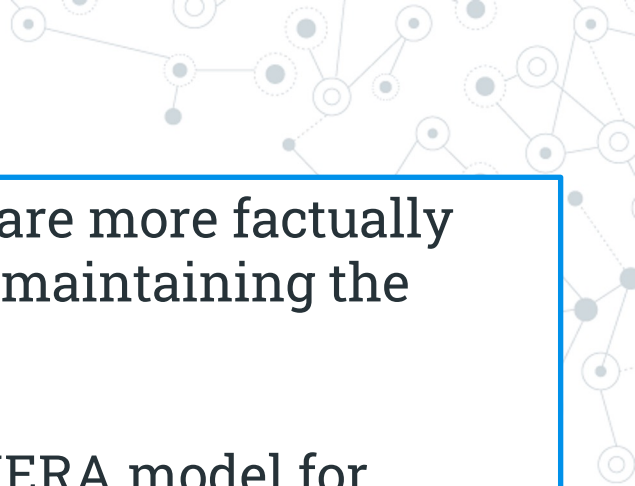

Trading in all Nasdaq-listed stocks and options was halted **for three hours on Thursday** due to technical problems on the bourse, MarketWatch reports. The exchange sent out a series of emails alerting investors that it was experiencing issues with "quote submissions." In response, the New York Stock Exchange has also stopped trading in all Nasdaq securities at the request of Nasdaq OMX. "All orders in those securities have been canceled back to customers," said NYSE in a statement. **The Nasdaq Composite Index was last at 3631.17, up 31.38 points, before trading was suspended.** There was no immediate word on when transactions will resume.

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by small circles, some of which are larger and have concentric circles, suggesting different levels of connectivity or importance. The lines are thin and gray, creating a mesh-like structure.

# CONCLUSIONS

A decorative network diagram in the bottom-right corner, similar to the one in the top-left. It shows a cluster of nodes connected by lines, with some nodes being larger and more prominent than others. The overall style is clean and modern, with a focus on geometric patterns.

- The problem addressed in the research paper is abstractive multi-document summarization (MDS) using cross-document graph-based information extraction (IE) .
- The goal is to generate a summary of a cluster of input documents by extracting the most salient information and reducing inconsistencies between the IE graphs and text representations.
- The paper proposes two novel components to enhance MDS performance: (1) the use of auxiliary entity and event recognition systems to focus the summary generation model, and (2) incorporating an alignment loss between IE nodes and their text spans.

- 
- The model aims to generate summaries that are more factually consistent with the source documents while maintaining the same level of abstractiveness.
  - The research also utilizes a pre-trained PRIMERA model for encoding the documents and a cross-entropy loss function for training the summary generation model.
- 



A photograph of three parallel strings of clear, round light bulbs hanging against a bright blue sky with soft, white clouds. The bulbs are slightly out of focus, creating a bokeh effect. The strings of lights run diagonally from the bottom left towards the top right.

**THANK YOU...**