# News Multi-document Summarization

Daniele Donia
d.donia@studenti.unisa.it
Matricola: 0522501575

## 1 Contesto del progetto

La grande quantità di testo e dati disponibili oggi, in particolare dal World Wide Web, presenta sfide significative per la comprensione e l'analisi. Di conseguenza, c'è una crescente necessità di tecniche di sintesi automatica, che facilitano l'estrazione di estratti concisi e rappresentativi da testi estesi. Tali tecniche hanno numerose applicazioni, tra cui il dominio delle notizie, dove aiutano a distillare grandi volumi di informazioni in brevi riassunti di notizie facilmente digeribili. Possiamo distinguere due approcci diversi per questo compito:

- Sintesi estrattiva, che prevede la selezione e il raggruppamento delle informazioni salienti da un testo sorgente per la creazione del riassunto.
- Sintesi astrattiva, che si basa sulla riformulazione delle idee principali del testo sorgente, anche includendo parole non presenti in origine.

Nel campo della sintesi di testo automatica, è necessario distinguere anche tra sintesi di un singolo documento (SDS) e di documenti multipli (MDS). Ma et al. [7] racchiude le differenze che caratterizzano la MDS nei seguenti aspetti:

- Diversità di tipo
- Relazioni tra i documenti
- Alta ridondanza e contraddizione nei documenti
- Spazio di ricerca più ampio

## 2 Obiettivi del progetto

L'obiettivo del progetto è di indagare il campo della sintesi automatica di documenti multipli, analizzando la qualità e la coesione dei riassunti generati. Ciò comprende l'approfondimento di diversi aspetti chiave di tale processo.

In primis, il progetto punta a determinare l'impatto dei vari step di pre-processing sulle performance e gli artifatti dei modelli di sintesi. Comprendendo come le fasi iniziali influenzano l'output, è possibile ottimizzare la qualità generale. Per questo motivo è definita la seguente domanda di ricerca:

> **RQ$_1$.** *In che misura le tecniche di pre-processing influenzano gli artefatti e le performance dei modelli di sintesi automatica?*

Si cerca, inoltre, di esplorare i potenziali benefici della combinazione di approcci estrattivi e astrattivi nel miglioramento della qualità dei riassunti. In particolare, si vuole attuare un confronto con l'utilizzo esclusivo di metodi astrattivi e estrattivi.

> **RQ$_2$.** *Può l'integrazione degli approcci estrattivi e astrattivi migliorare la qualità dei riassunti rispetto ai singoli approcci?*

Attraverso tali domande di ricerca, il progetto ha l'obiettivo di sviluppare un sistema di sintesi multi-documento in grado di fornire, dato un insieme di articoli giornalistici in input, un riassunto coinciso e informativo.

## 3 Step metolodologi per la realizzazione degli obiettivi

### 3.1 Contesto dello studio

All'inizio è stata condotta un'analisi dello stato dell'arte relativa alla sintesi multi-documento. Tale analisi include la comprensione degli studi precedenti e dei dataset impiegati in modo da avere una visione completa di ciò che è stato raggiunto dalle tecnologie attualmente utilizzate. I dataset di larga scala per questo tipo di task sono relativamente scarsi [7]; in particolare, i più rappresentativi nell'ambito delle notizie giornalistiche sono descritti di seguito:

- **DUC e TAC**: i dataset DUC (Document Understanding Conference) e TAC (Text Analysis Conference) sono stati creati per promuovere la ricerca nel campo della sintesi automatica di testi, fornendo competizioni ufficiali dal 2001 al 2007. Questi dataset sono composti principalmente da articoli di cronaca provenienti da varie categorie tematiche, tra cui politica, disastri naturali e biografie. I dataset DUC e TAC offrono un numero limitato di documenti, contenendo solo alcune centinaia di articoli di notizie e riassunti annotati manualmente, utili per la valutazione delle prestazioni dei modelli di sintesi.
- **Wikisum**: è un dataset per la sintesi automatica, sviluppato a partire dalla base di conoscenza di WikiHow. I documenti inclusi nel dataset sono redatti in un inglese semplice e le sintesi sono presentate sotto forma di paragrafi coerenti, redatti direttamente dagli autori dei testi originali. Come evidenziato da Cohen et al. [3], le principali caratteristiche di WikiSum sono quindi: sintesi organizzate in un singolo paragrafo coerente, articoli e sintesi di facile lettura e ridotta necessità di conoscenze pregresse per una piena comprensione.
- **Multinews**: è un insieme di dati di ampia scala nel dominio delle notizie, composto da articoli e riassunti redatti manualmente, tutti provenienti dal web. Questo dataset comprende 56.216 coppie di articoli e riassunti e include link di riferimento ai documenti originali. Inoltre, Fabbri et al. [4] hanno confrontato il dataset Multi-News con i dataset precedenti in termini di copertura, densità e compressione, evidenziando che il Multi-News presenta diverse modalità di disposizione delle sequenze.

### 3.2 Stato dell'arte

Per loro stessa natura, i molteplici documenti di input utilizzati nella sintesi automatica di testi multipli (MDS, Multi-Document Summarization) tendono a contenere informazioni contraddittorie, ridondanti e complementari. Pertanto, i modelli MDS richiedono

algoritmi sofisticati per identificare e gestire la ridondanza e le contraddizioni tra i documenti, al fine di garantire che il riassunto finale sia completo. Due metodi comuni per concatenare più documenti sono i seguenti:

- **Concatenazione piatta**: un metodo di concatenazione semplice ma efficace, in cui tutti i documenti di input vengono trattati come una sequenza piatta. In una certa misura, questo metodo converte il compito di MDS in un compito di sintesi di singolo documento (SDS, Single-Document Summarization).
- **Concatenazione gerarchica**: questo metodo è in grado di preservare le relazioni tra i documenti. Tuttavia, molti metodi di deep learning esistenti non sfruttano appieno questa relazione gerarchica. Sfruttare le relazioni gerarchiche tra i documenti, invece di concatenarli semplicemente in modo piatto, permette al modello MDS di ottenere rappresentazioni con informazioni gerarchiche integrate, migliorando così l'efficacia dei modelli. Gli algoritmi di clustering e le tecniche basate su grafi sono i metodi più comunemente utilizzati.

La ricerca nel campo della sintesi automatica di testi ha avuto inizio con la sintesi di singoli documenti (SDS, Single-Document Summarization) per poi passare successivamente ai documenti multipli (MDS, Multi-Document Summarization). Negli ultimi anni, sono stati proposti diversi approcci per la MDS. Afsharizadeh et al. [1] hanno categorizzato i modelli di MDS in sei categorie principali:

- **Machine Learning**: Questa categoria comprende tecniche basate su algoritmi di apprendimento supervisionato e non supervisionato, come le Support Vector Machines (SVM), le reti neurali artificiali (ANN) e gli alberi di decisione. Tali metodi vengono utilizzati per imparare schemi rilevanti nei testi, quali le strutture sintattiche e semantiche. Ad esempio, le SVM possono essere impiegate per classificare frasi come pertinenti o meno per un riassunto
- **Clustering**: tali tecniche, come il k-means e l'analisi delle componenti principali (PCA), sono utilizzate per raggruppare frasi o paragrafi con contenuti simili, identificando diversi argomenti trattati nei testi di input. Questo processo facilita l'individuazione delle sezioni centrali dei documenti e riduce la ridondanza selezionando una rappresentazione centrale di ogni cluster per il riassunto finale.
- **Metodi basati su grafi**: tali metodi rappresentano i documenti come grafi in cui i nodi corrispondono a frasi o paragrafi e gli archi rappresentano relazioni semantiche o di similarità tra di essi. Tecniche basati su grafi: algoritmi come TextRank e LexRank utilizzano rappresentazioni a grafo dei documenti per identificare le frasi più centrali e informative.
- **Approcci basati su LDA (Latent Dirichlet Allocation)**: LDA è un modello generativo che rappresenta i documenti come combinazioni di argomenti, ciascuno dei quali è rappresentato da una distribuzione di parole. Attraverso metodi bayesiani e di massimizzazione dell'aspettativa (EM, Expectation-Maximization), LDA può identificare argomenti e termini chiave all'interno di un insieme di documenti. Questo approccio è efficace per riassumere documenti in cui gli argomenti sono distribuiti in modo disomogeneo o si sovrappongono.
- **Deep Learning**: le reti neurali profonde, in particolare i modelli basati su Transformer, hanno rivoluzionato il campo della sintesi automatica di testi. Questi modelli sono in grado di processare grandi quantità di testo e di apprendere rappresentazioni multiple e gerarchiche delle sequenze di input.

Gli approcci correnti alla MDS si possono suddividere ulteriormente in tre categorie: estrattivi, astrattivi e ibridi.

I metodi estrattivi [6] selezionano e combinano frasi dai documenti di origine. Questi approcci includono:

- Metodi basati su ontologie: utilizzo di ontologie specifiche di dominio per identificare concetti chiave e selezionare frasi rilevanti (ne è un esempio YAGO Summarizer proposto da Baralis et al. [2]).
- Tecniche basate su grafi
- Metodi basati sui termini (che possono essere ulteriormente classificati in metodi di clustering, metodi di analisi semantica latente e fattorizzazione non negativa di matrici)
- Metodi basato sulla teoria della struttura retorica: suddividono il testo in unità testuali adiacenti, come frasi consecutive, e applicano diverse regole RST per valutare l'importanza di ciascuna unità. Le frasi vengono quindi classificate in nuclei e satelliti (che forniscono informazioni aggiuntive sui nuclei).

I metodi estrattivi sono apprezzati per la loro fedeltà al testo originale e l'efficienza computazionale. Tuttavia, possono produrre riassunti meno fluidi e sono limitati dalle frasi presenti nei documenti originali.

Gli approcci astrattivi generano nuovo testo catturando l'essenza dei documenti originali. Questi si basano su architetture avanzate come:

- Modelli basati su trasformers: modelli pre-addestrati come BART, T5 o PEGASUS hanno dimostrato un'eccezionale performance nella generazione di riassunti coerenti e informativi.
- Modelli basati su apprendimento per rinforzo, utilizzato per ottimizzare i risultati sulla base della valutazione umana o metriche di valutazione come ROUGE.

Gli approcci astrattivi offrono maggiore flessibilità e potenziale creativo, generando riassunti più concisi e coerenti. Tuttavia, affrontano sfide come il rischio di "allucinazioni" (generazione di informazioni non presenti nei documenti originali) e una maggiore intensità computazionale [13].

Gli approcci ibridi combinano elementi estrattivi e astrattivi per sfruttare i vantaggi di entrambi. Ne è un esempio **HMSumm** (definito da Ghadimi and Beigy [5]), in cui si utilizza BERT per rappresentare contestualmente le frasi, per poi creare un grafo basato sulla loro similarità. Le frasi sono valutate tramite un modello Determinantal Point Process (DPP) con DSN per selezionare le più importanti. Queste frasi costituiscono un riassunto estrattivo, che viene poi elaborato dai modelli pre-addestrati BART e T5 per la generazione del riassunto.

*3.2.1 Metriche di valutazione* La valutazione della qualità dei riassunti generati automaticamente rappresenta una sfida complessa nel campo del Natural Language Processing (NLP), richiedendo una comprensione approfondita delle diverse metriche di valutazione

disponibili. Queste metriche possono essere classificate in due categorie principali: intrinseche ed estrinseche.

Le metriche di valutazione intrinseche sono progettate per misurare la qualità dei riassunti basandosi esclusivamente sul loro contenuto e struttura, senza considerare fattori esterni.
ROUGE è una delle metriche intrinseche più ampiamente utilizzate. Misura la sovrapposizione tra il riassunto generato e un insieme di riassunti di riferimento, confrontando n-grammi o sequenze di parole.

- **ROUGE-N**: Valuta la sovrapposizione degli n-grammi. È particolarmente apprezzata per la sua semplicità ed efficacia nella valutazione della copertura del contenuto.
- **ROUGE-L**: Valuta la più lunga sottosequenza comune, fornendo intuizioni sulla fluidità e coerenza del riassunto.

Nonostante il suo ampio utilizzo, ROUGE presenta alcune limitazioni, tra cui la dipendenza dal match esatto e la sensibilità alla lunghezza dei riassunti.
Per superare le limitazioni degli approcci basati sul match esatto, sono state sviluppate metriche di valutazione semantica. Ne è un esempio **BERTScore**, una metrica che utilizza le rappresentazioni contestuali di BERT (Bidirectional Encoder Representations from Transformers) per calcolare i punteggi di somiglianza tra il riassunto generato e i riassunti di riferimento. BERTScore misura la precisione, il richiamo e il punteggio F1 basandosi sulla somiglianza a livello di token, offrendo una valutazione più robusta e semanticamente informata.

Le metriche estrinseche valutano l'impatto dei riassunti su compiti successivi o sulla soddisfazione dell'utente, fornendo una prospettiva più ampia sull'utilità pratica dei riassunti generati.

Le valutazioni umane rimangono un aspetto essenziale nella valutazione della qualità della sintesi. I giudici umani possono fornire intuizioni qualitative su aspetti come:

- Leggibilità
- Coerenza
- Informatività

Questi aspetti potrebbero non essere completamente catturati dalle metriche automatizzate. Framework di valutazione strutturati, come il metodo **Pyramid**, coinvolgono più valutatori umani che valutano l'importanza delle informazioni nei riassunti generati rispetto ai riassunti di riferimento.

Insieme, queste metriche forniscono un framework completo per valutare la qualità della generazione dei riassunti, guidare i miglioramenti delle prestazioni del modello e garantire che i riassunti soddisfino le esigenze di applicazioni e utenti diversi.

## 3.3 Analisi dei dati

Il dataset Multi-News, una raccolta di articoli di notizie provenienti da varie fonti e aggregati per la sintesi multi-documento, è stato analizzato per comprenderne le proprietà strutturali.

*3.3.1* ***Preparazione dei Dati e Panoramica*** Il dataset è stato caricato utilizzando la libreria `datasets` di Hugging Face e poi diviso in insiemi di addestramento, validazione e test. Ogni sottoinsieme è stato convertito in DataFrames di pandas, che sono stati successivamente concatenati in un unico DataFrame per facilitare un approccio di analisi unificato.

*3.3.2* ***Risultati***

(1) **Comprensione del dataset:** Il dataset è composto da due caratteristiche: 'document' e 'summary'. Contiene 56216 istanze e non sono presenti né valori nulli né valori duplicati.
(2) **Distribuzione della lunghezza dei documenti:** L'analisi delle lunghezze dei documenti ha rivelato una significativa variabilità all'interno del dataset. Le lunghezze dei documenti, misurate in termini di conteggio delle parole, mostrano un ampio intervallo. La mediana della lunghezza dei documenti è risultata essere di circa **1.540 parole**, con il documento più corto contenente circa **0 parole** e il più lungo che arriva fino a **519.721 parole**. Questo ampio intervallo suggerisce che il dataset include sia brevi notizie che articoli più estesi, il che potrebbe rappresentare una sfida per i modelli di sintesi che devono gestire efficacemente diverse dimensioni di contenuto.
La distribuzione delle lunghezze dei documenti sembra essere asimmetrica a destra (Fig.1), indicando una maggiore frequenza di documenti più brevi e una lunga coda di articoli più lunghi.
(3) **Distribuzione della lunghezza dei riassunti:** I riassunti presenti nel dataset sono stati analizzati per valutarne la distribuzione delle lunghezze. La mediana della lunghezza è di circa **260 parole**, con il sommario più breve di **41 parole** e il più lungo che raggiunge le **1199 parole**.
La distribuzione delle lunghezze dei sommari (Figura 2) è più uniforme rispetto a quella dei documenti, suggerendo che i sommari tendono ad aderire a una struttura e lunghezza più coerenti, indipendentemente dalla lunghezza originale del documento.
(4) **Analisi della Frequenza delle Parole:** Le parole più frequenti nel dataset includono termini tipici delle notizie come mostrato nell'istogramma sottostante (DA INSERIRE).Anche le parole di uso comune come "il," "e," "di," "a," e "in" dominano il dataset. Questo risultato sottolinea l'importanza di passaggi di pre-elaborazione come la rimozione delle parole di uso comune quando si prepara il dato per i modelli di apprendimento automatico.
(5) **Analisi della Diversità delle Fonti:** Il dataset è notevole per la sua diversità di fonti, con ogni documento che tipicamente cita più fonti di notizie. In media, ogni articolo fa riferimento a circa 2 fonti diverse (fino a un massimo di 10), riflettendo una gamma di prospettive.
La varietà nelle fonti introduce anche sfide legate al bilanciamento delle diverse prospettive e all'assicurare che i modelli non favoriscano sproporzionatamente fonti specifiche nei loro sommari.

# 4 Metodologia

In relazione alla domanda di ricerca $\mathbf{RQ_1}$., sono state considerate le seguenti tecniche di pre-processing:

- **Rimozione delle stop word**: questo metodo elimina parole di alta frequenza e basso valore informativo, come articoli, preposizioni, congiunzioni e pronomi, che non apportano un contributo sostanziale alla rappresentazione semantica del testo.

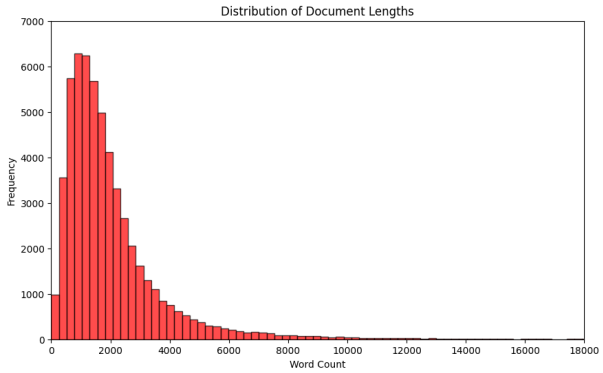


**Figure 1: Distribuzione lunghezza documenti**

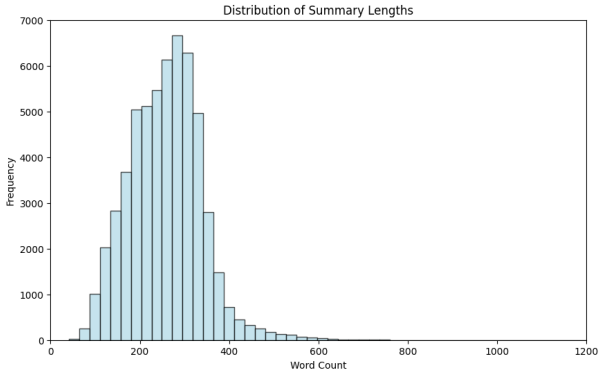


**Figure 2: Distribuzione lunghezza riassunti**

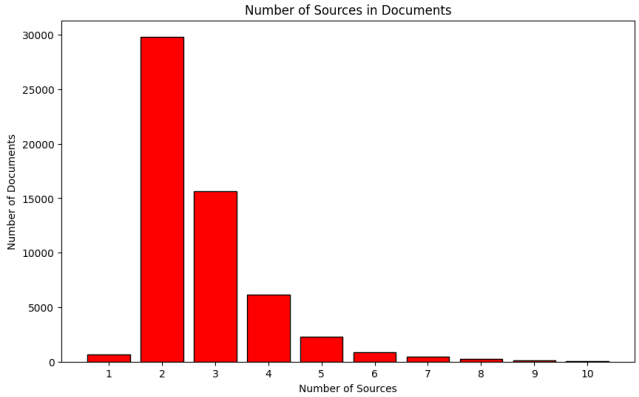


**Figure 3: Distribuzione numero fonti**

- **Stemming**: una tecnica di normalizzazione linguistica che riduce le forme flesse o derivate delle parole alla loro radice (stem), al fine di diminuire la variabilità lessicale. Lo stemming si basa su regole euristiche per produrre una rappresentazione compatta, sebbene questa possa non corrispondere a una parola grammaticalmente corretta.

Nell'ambito dell'analisi è stato progettato un modello ibrido per la sintesi multi-documento, integrando un modulo estrattivo con un modulo astrattivo. Questa architettura sfrutta i vantaggi di entrambi gli approcci: il modulo estrattivo identifica le informazioni chiave dai documenti di input, mentre il modulo astrattivo genera un riassunto più fluido e semantico.

Per il modulo estrattivo, sono state esplorate due tecniche:

- **TextRank**: un algoritmo non supervisionato basato su grafi. Nello specifico per ciascun documento si costruisce un grafo in cui i nodi rappresentano frasi chiave, e gli archi tra i nodi indicano la similarità tra di essi. La similarità tra due frasi è misurata tramite la distanza coseno basata su vettori di rappresentazione. L'algoritmo assegna un punteggio ad ogni frase del documento, calcolato iterativamente in base alla centralità della frase rispetto alle altre nel grafo. Frasi con punteggi elevati vengono considerate più importanti e rappresentative del contenuto del documento.
- **K-means clustering**: una tecnica di raggruppamento non supervisionato che segmenta le frasi in cluster, selezionando rappresentanti centrali di ciascun cluster come le frasi più rilevanti. In particolare, ciascun documento in input è dato in input a un encoder BERT per ottenere una rappresentazione contestualizzata. Tale rappresentazione è poi utilizzata nell'algoritmo di clustering per determinare le frasi più vicine ai centroidi di ciascun cluster, ovvero quelle più rappresentative.

Le frasi estratte da ciascun documento vengono poi concatenate e fornite come input al modulo astrattivo, che si avvale di modelli basati su *Transformers*, capaci di comprendere contesti complessi e produrre testi coerenti e sintetizzati. Per la componente astrattiva del sistema, sono state valutate due architetture basate su Transformer:

- T5-small: Una variante ridotta del modello T5 (Text-to-Text Transfer Transformer) sviluppato da Raffel et al. (2020). T5 adotta un approccio unificato ai compiti di NLP, trattando ogni attività come una trasformazione da testo a testo. T5-small contiene circa 60 milioni di parametri, significativamente meno rispetto ai 220 milioni della versione base, offrendo un compromesso tra efficienza computazionale e capacità di generalizzazione.
- Distill-BART: Una versione distillata di BART (Bidirectional and Auto-Regressive Transformers) proposta da Lewis et al. (2020). BART combina un encoder bidirezionale con un decoder autoregressivo. La versione distillata mantiene gran parte delle capacità del modello originale, ottimizzando al contempo le prestazioni in termini di velocità di esecuzione.

T5-small offre maggiore flessibilità e leggerezza, rendendolo adatto per applicazioni con risorse computazionali limitate o scenari multi-task. Tuttavia, la sua capacità di generazione testuale potrebbe essere inferiore rispetto a modelli più ampi, specialmente in compiti di sintesi complessi. Distill-BART, grazie alla sua architettura che combina codifica bidirezionale e decodifica autoregressiva, potrebbe

fornire una qualità di sintesi superiore. Sebbene sia computazionalmente più intensivo di T5-small, rappresenta un equilibrio tra efficienza e prestazioni elevate per task di generazione testuale e sintesi.

Il modulo astrattivo è responsabile della produzione del riassunto finale, ottenendo una sintesi coerente e concisa dell'intero set di documenti, migliorando così sia la leggibilità che la qualità del riassunto rispetto a metodi puramente estrattivi.

## 5 Implications of the results

## 6 Conclusion

## References


[1] Mahsa Afsharizadeh, Hossein Ebrahimpour-Komleh, Ayoub Bagheri, and Grzegorz Chrupała. 2022. A survey on multi-document summarization and domain-oriented approaches. *Journal of Information Systems and Telecommunication (JIST)* 1, 37 (2022), 68.

[2] Elena Baralis, Luca Cagliero, Saima Jabeen, Alessandro Fiori, and Sajid Shah. 2013. Multi-document summarization based on the Yago ontology. *Expert Systems with Applications* 40, 17 (2013), 6976–6984.

[3] Nachshon Cohen, Oren Kalinsky, Yftah Ziser, and Alessandro Moschitti. 2021. Wikisum: Coherent summarization dataset for efficient human-evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 212–219.

[4] Alexander R Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv:1906.01749* (2019).

[5] Alireza Ghadimi and Hamid Beigy. 2022. Hybrid multi-document summarization using pre-trained language models. *Expert Systems with Applications* 192 (2022), 116292.

[6] Zakia Jalil, Jamal Abdul Nasir, and Muhammad Nasir. 2021. Extractive multi-document summarization: a review of progress in the last decade. *IEEE Access* 9 (2021), 130928–130946.

[7] Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z Sheng. 2022. Multi-document summarization via deep learning techniques: A survey. *Comput. Surveys* 55, 5 (2022), 1–37.