Exploring the Impact of Extractive Summarization on GPT-based Abstractive Summarization

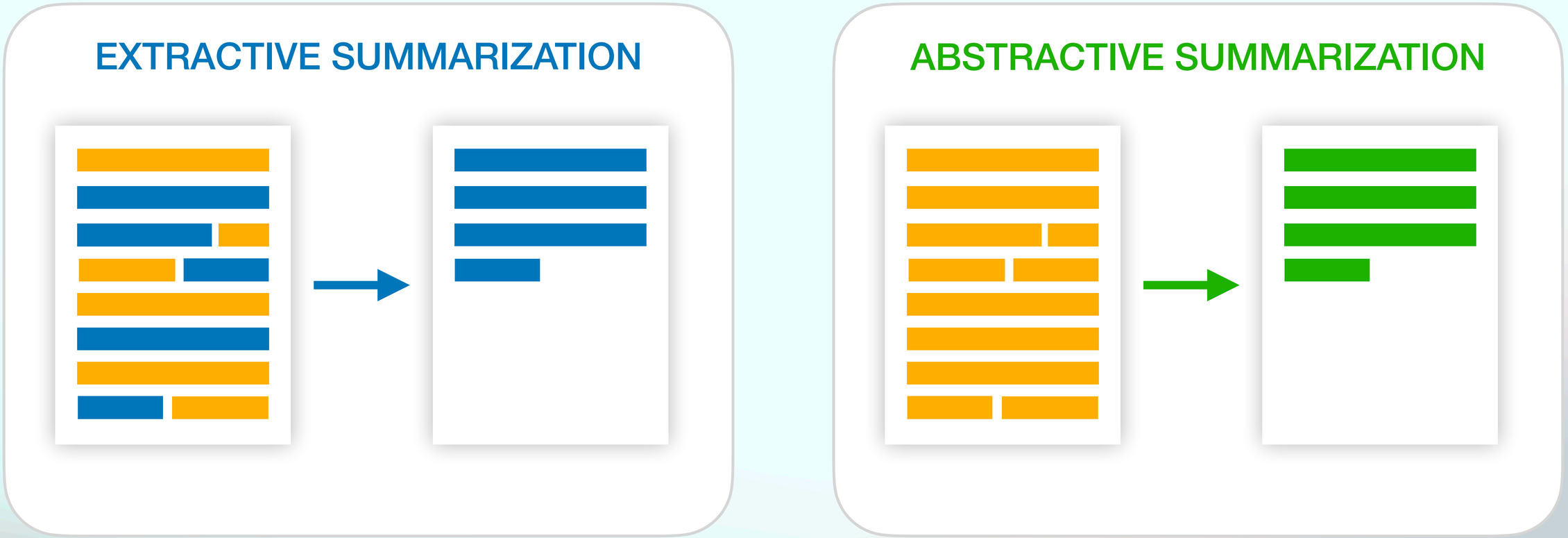
Corso di Laurea Magistrale in Ingegneria Informatica

ANNO ACCADEMICO 2023/2024

Candidato: **Simone Giorgio**

Relatore: **Prof. Andrea Tagarelli**

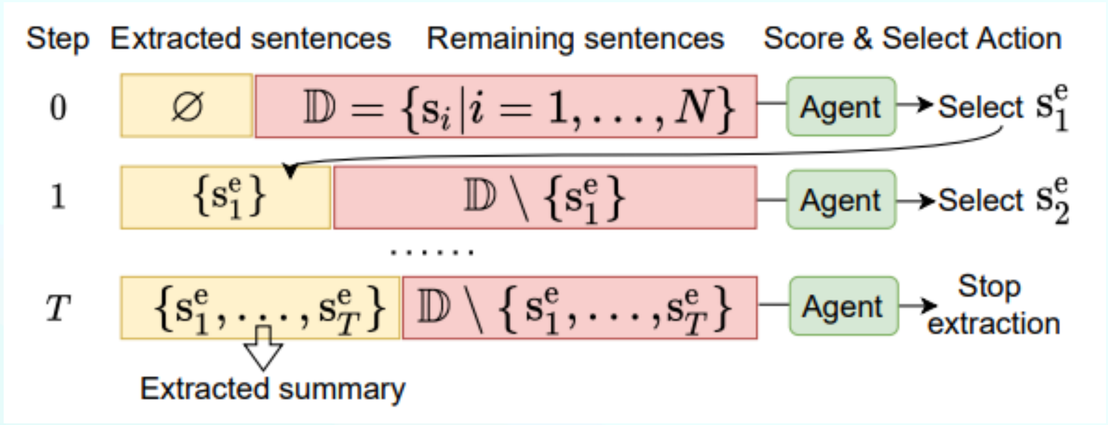
Introduzione



- Studio MemSum e KeyBERT
- Valutazione impatto extractive summarization su GPT-3.5

MemSum - Multi-step Episodic Markov decision process extractive Summarizer

Modello Reinforcement Learning based per la extractive summarization, Nianlong Gu et al. ACL 2022



Agent aggiorna ad ogni timestep l'informazione sull'extraction history corrente prima di selezionare prossima azione

Policy Gradient Methods

Tecniche RL basate su ottimizzazione policy parametrizzata mediante gradient descent

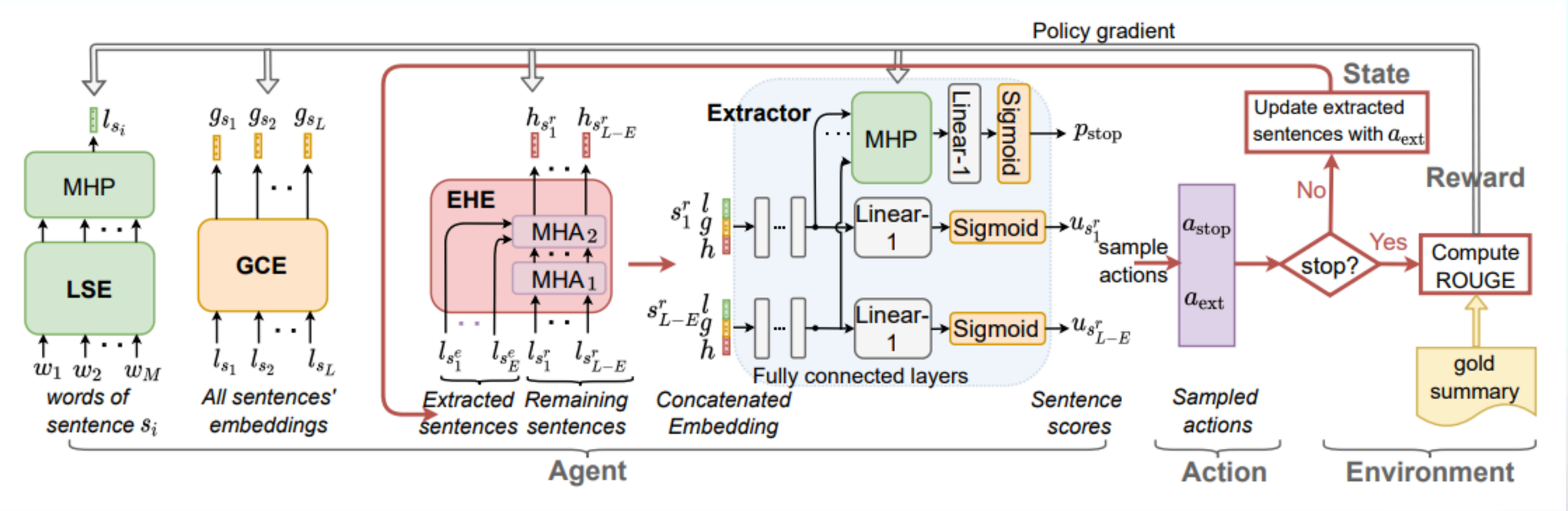
∀ timestep ***SENTENCE STATE*** informativo su

- Sentence Local Content
- Sentence Global Context nel documento
- Info su extraction history

↓

AGENT → **SCORES**

MemSum



Per codificare le 3 proprietà nello stato si usano:

Local Sentence Encoder

2 layers bi-LSTM trasformano word embeddings in sentence embeddings con MHP

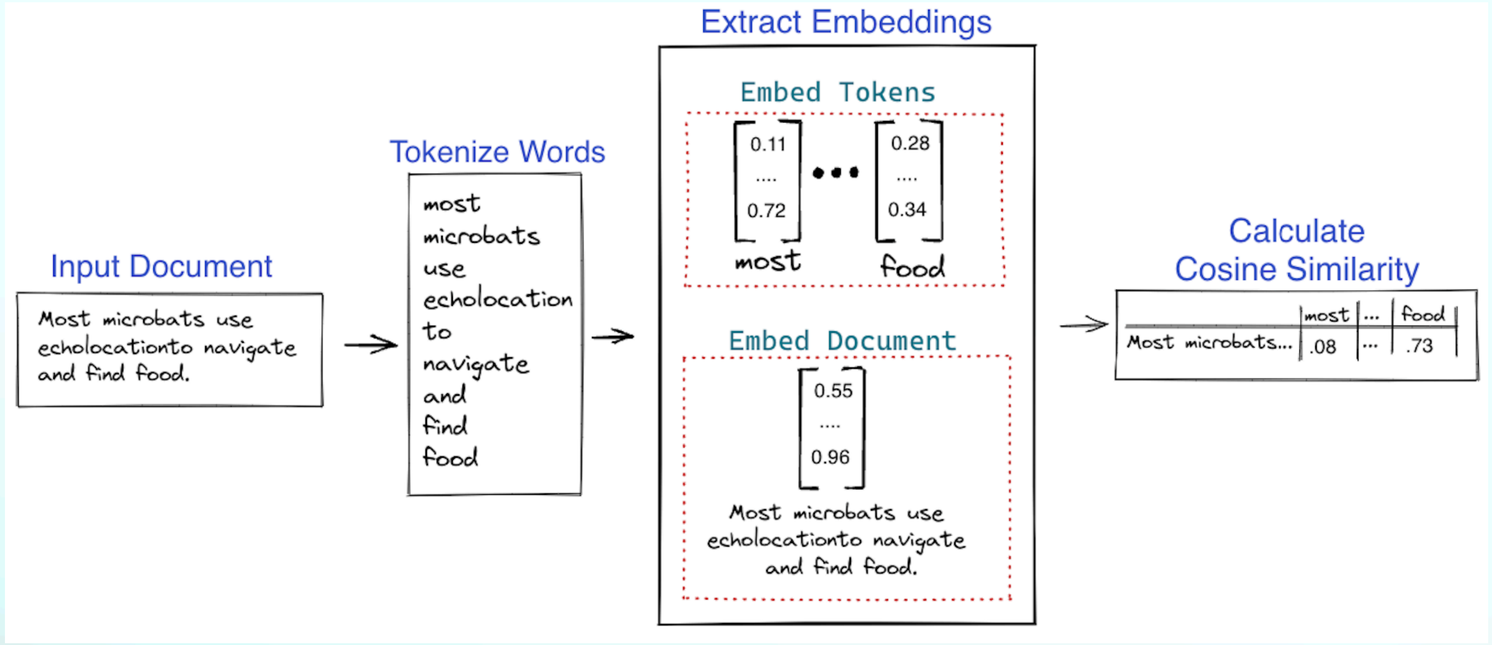
Global Context Encoder

2 layers bi-LSTM producono a partire dai sentence embeddings un global embedding per ogni sentence

Extraction History Encoder

3 Attention layers, ognuno composto da 2 MHA sublayers, produce history embeddings per ogni sentence rimanente

Key-BERT for Keywords extraction



- **BERT-based embeddings** per ottenere rappresentazioni a livello di singoli tokens e di documento
- **Similarità del coseno** per individuare i segmenti di testo più simili al documento stesso

Sentence-Transformers:

framework per embeddings che offre ampia selezione di pre-trained models tuned per diversi tasks.

Pre-trained model utilizzato ***all-mpnet-base-v2*** *:*

general purpose, addestrato su un dataset vasto e diversificato formato da oltre 1 miliardi di elementi.
Limite 384 tokens.

Criteri di Valutazione

ROUGE scores

Valutano similarità tra summary generato e riferimento,
in termini di soprapposizione di unità di testo

- **ROUGE-1:** misura sovrapposizione uni-grams
- **ROUGE-2:** misura sovrapposizione bi-grams
- **ROUGE-Lsum:** Misura la lunghezza della sequenza di parole in comune che appaiono nello stesso ordine *(LCS, Longest Common Subsequence)*

espressi in termini di

Precision: $\frac{\text{Numero di parole comuni tra i summaries}}{\text{Numero di parole nel summary di riferimento}}$

Recall: $\frac{\text{Numero di parole comuni tra i summaries}}{\text{Numero di parole nel summary generato}}$

F1-score: $2 \times \frac{P \times R}{P + R}$

Sistema di Ratings

Richiesta inviata a GPT-3.5 :

"Given the following summary of the source text rate it on a scale 1(worst)-5(best) according to five criteria:

- [1] **Informative :** a summary is informative if it encapsulates the crucial details from the source, offering a precise and concise presentation.
- [2] **Quality :** a summary has an high quality if it is understandable and comprehensible.
- [3] **Coherence :** a summary is coherent if it demonstrates a sound structure and organization.
- [4] **Attributable :** all the information in the summary are attributable to the source.
- [5] **Overall Preference :** the summary should succinctly, logically, and coherently convey the primary ideas presented in the source."

Datasets Utilizzati

Oggetto dello studio 9 datasets:

1. **Arxiv**: Pre-print di articoli scientifici in fisica, matematica, informatica, statistica, finanza quantitativa e biologia;
2. **GovReport**: Relazioni scritte da agenzie di ricerca governative Statunitensi;
3. **PubMed**: Letteratura scientifica biomedica;
4. **EurLexSum**: Atti legali emessi dall'unione europea;
5. **BigPatent**: Documentazione di brevetti Statunitensi nell'ambito delle "human necessities";
6. **eLife**: Articoli scientifici dall'eLife Journal;
7. **BookSum**: Documenti nel dominio della letteratura, come romanzi, opere teatrali e storie;
8. **Multi-News**: Articoli da newser.com;
9. **WikiSum**: Tutorial su svariati argomenti da WikiHow.

Preprocessing

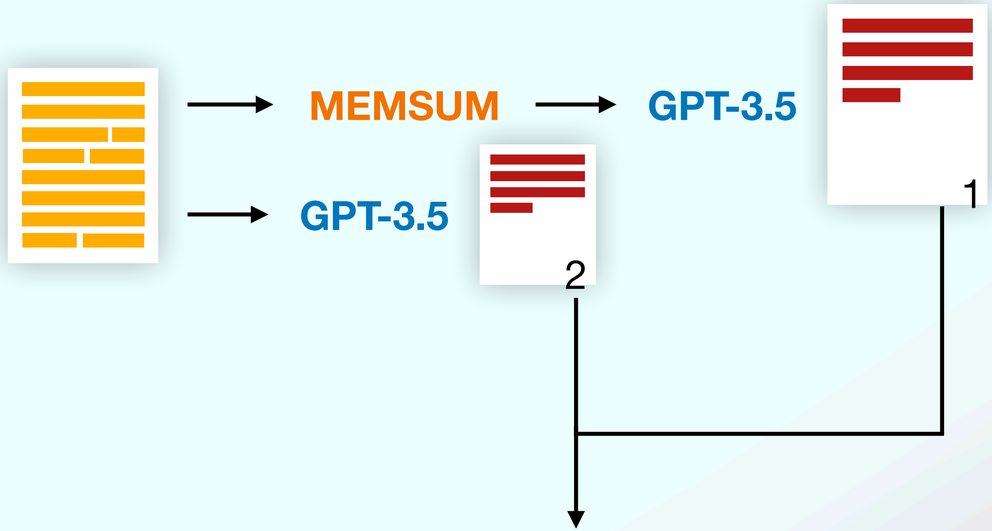
- Eliminazione features superflue e modifica per adattare a formato [*Document - Summary*]
- Trasformazione testo in lista sentences utilizzando *sentences_tokenize* di nltk

Training MemSum Models e prima pipeline con GPT-3.5

TRAINING MODELLI CON 250 DOCUMENTI

Memsum Trained On	rouge1	rouge2	rougeLsum
Arxiv	0.479	0.199	0.421
PubMed	0.493	0.230	0.444
GovReport	0.594	0.285	0.567
EurLexSum	0.478	0.231	0.468
BigPatent	0.423	0.215	0.376
eLife	0.438	0.094	0.407
BookSum	0.286	0.056	0.274
MultiNews	0.431	0.140	0.391
WikiSum	0.306	0.071	0.277

PRIMA PIPELINE MEMSUM - GPT-3.5



“Quale pensi sia migliore riassunto tra i seguenti?”

Il secondo riassunto è più chiaro e conciso rispetto al primo. Riassume le principali disposizioni della regolazione in modo più accessibile e fornisce una visione più chiara delle sue implicazioni. Tuttavia, entrambi i riassunti includono informazioni importanti sulla regolazione, quindi la scelta tra i due dipenderà dal contesto e dall'audience a cui è destinato il riassunto. Se si tratta di un pubblico tecnico o legale, il primo riassunto potrebbe essere più appropriato, ma se si vuole comunicare le informazioni in modo più accessibile, il secondo riassunto è preferibile.

ESEMPIO FUNZIONE ESTRAZIONE

```
extracted_summary_batch, extracted_indices_batch = memsum_custom_data.extract( [ document ],
                                    p_stop_thres = 0.6,
                                    max_extracted_sentences_per_document = 28,
                                    return_sentence_position=1
                                   )
```

ESEMPIO FUNZIONE VALUTAZIONE

```
[ ]  evaluate( memsum_custom_data, Fulltest, 0.6, 28, rouge_cal )

     100%|██████████| 188/188 [04:36<00:00,  1.47s/it]
     array([0.4778956 , 0.23138885, 0.46175933])
```

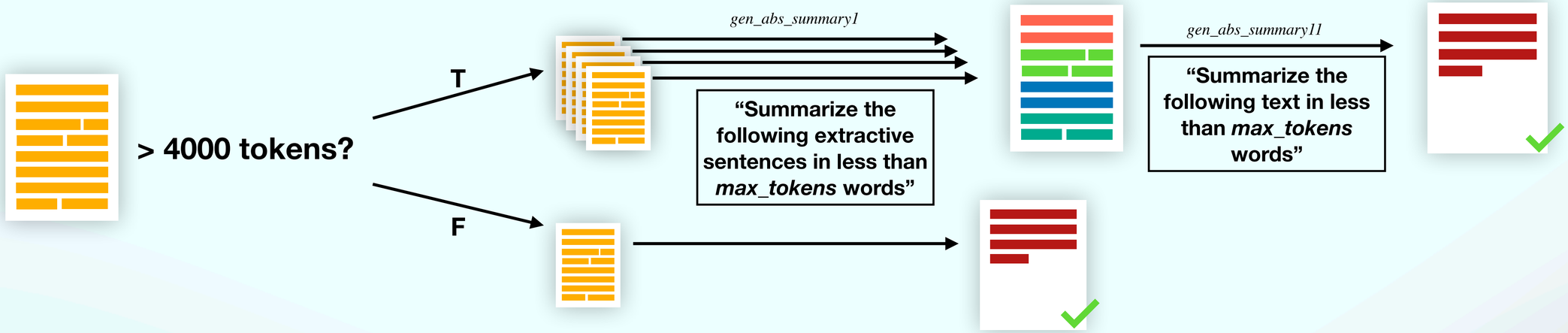
Summarization con assistenza estrattiva mantiene specificità e dettaglio tecnico maggiore.

Pipeline Finale

OBIETTIVO:
VALUTARE IMPATTO
ASSISTENZA ESTRATTIVA
SU GPT-3.5

4 MODALITÀ DI GENERAZIONE

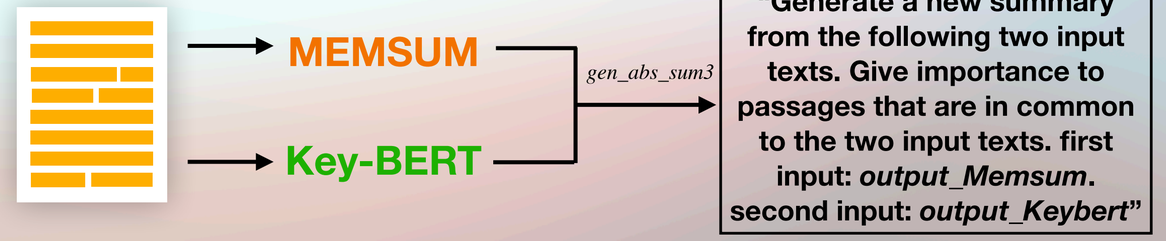
MODE 0: Nessun condizionamento, GPT-3.5 puro.



MODE 1: Condizionamento GPT-3.5 da parte di MemSum.

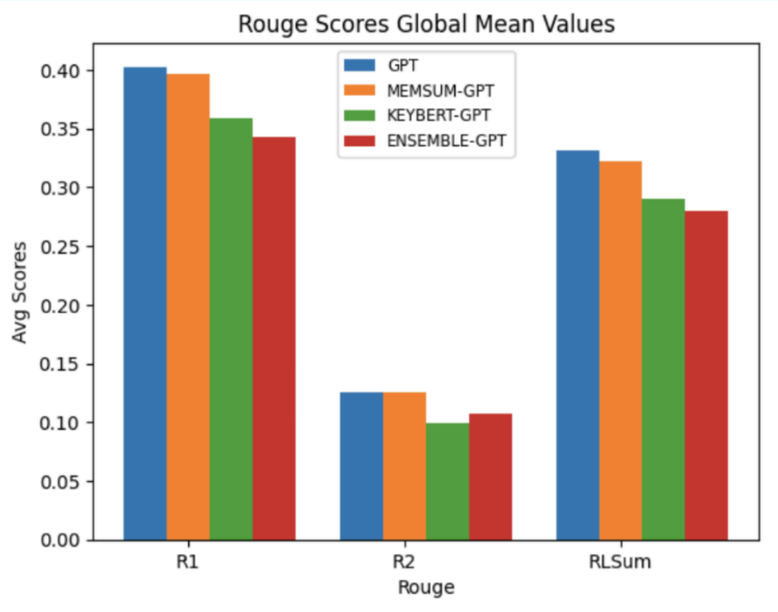
MODE 2: Condizionamento GPT-3.5 da parte di Key-BERT.

MODE 3: Condizionamento GPT-3.5 da parte di ensemble MemSum-KeyBERT



Risultati

CALCOLATI SU CAMPIONE 10 DOCS PER DATASET



COMPRESSION RATE MEDIO

	MODE0	MODE1	MODE3	MODE4
0	0.919937	0.943355	0.944966	0.962249
1	0.903484	0.925843	0.920809	0.959041
2	0.937941	0.950729	0.947875	0.964444
3	0.876157	0.929738	0.929130	0.956621
4	0.928887	0.944753	0.938622	0.970139
5	0.965598	0.965509	0.968891	0.970468
6	0.950777	0.943553	0.955430	0.970721
7	0.945191	0.949895	0.948764	0.958045
8	0.932151	0.948216	0.953526	0.968117
Media	0.928903	0.944621	0.945335	0.964427

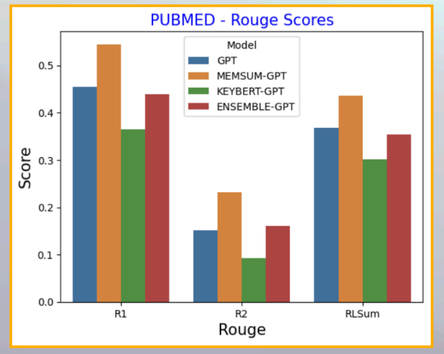
SUMMARIES OTTENUTI
CON ASSISTENZA ESTRATTIVA
MEDIAMENTE PIU' COMPATTI ✓

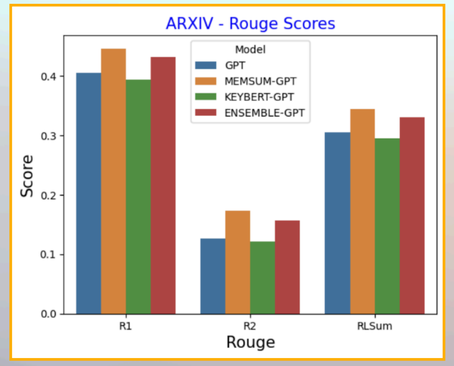
RATINGS MEDI GLOBALI ENSEMBLE-GPT:

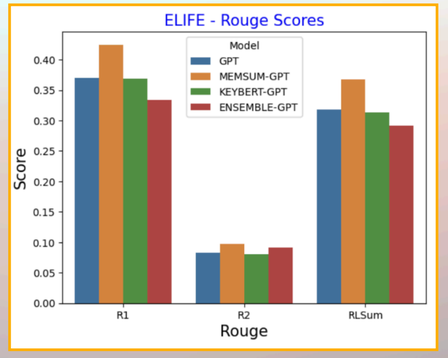
Informative	4,4
Quality	4,3
Coherence	4,2
Attributable	4,5
Overall Preference	**4,3**

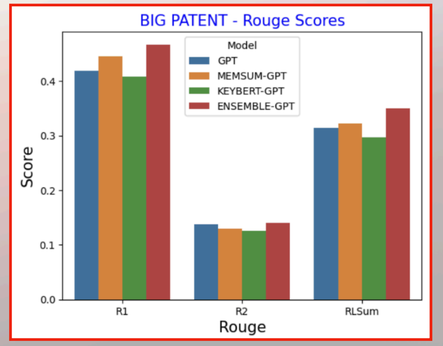
SUMMARIES DI OTTIMA QUALITÀ ✓

RISULTATI MIGLIORI CON ASSISTENZA ESTRATTIVA SU SINGOLI DATASET:





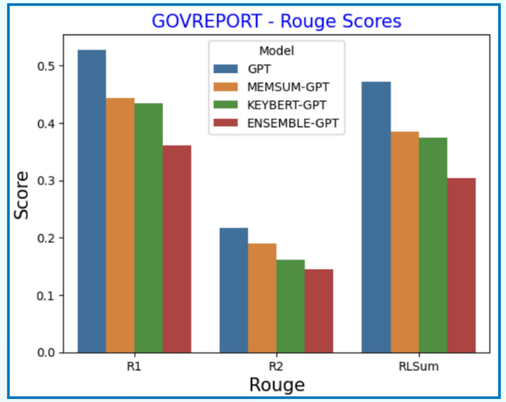




Analisi Finale

MODE0: [-38%, -43%, -45%]

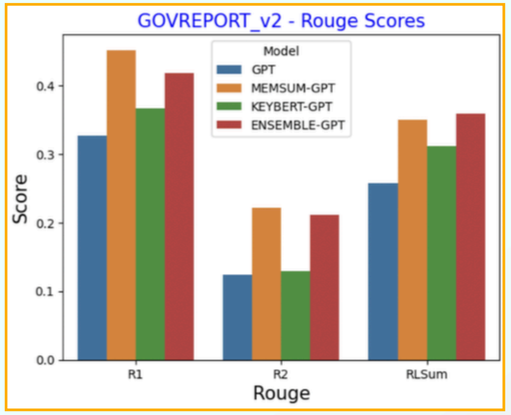
MODE3: **[+16%, +45%, +18%]**



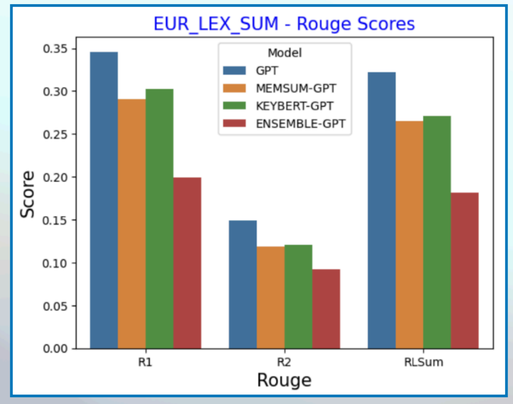
C-R mediano = 0.15

$$\textit{Compression-Ratio} = \frac{\text{numero words summary}}{\text{numero words documento}}$$

Datasets caratterizzati da Compression-Ratio più elevato sono quelli in cui le modalità con assistenza estrattiva hanno ottenuto gli scores minori.



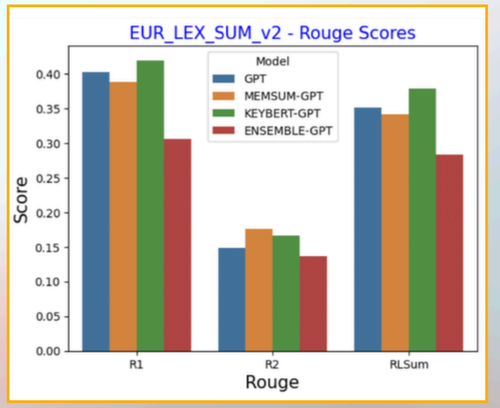
C-R mediano = 0.027



C-R mediano = 0.29

MODE3 migliore su BIGPATENT :
Compression-Ratio mediano=0.026

Dopo aver appositamente selezionato i docs con i valori di C-R minori



C-R mediano = 0.14

MODE0: [+16%, -0.5%, +9%]

MODE3: **[+53%, +49%, +56%]**

Considerazioni finali

- Fattori che possono influenzare i ROUGE scores: natura e qualità summary riferimento, numero sentences estratte e numero di tokens finali da far generare al modello astrattivo, nonché la *temperatura* della risposta.

- Verificato tramite il meccanismo di Ratings introdotto che i riassunti ottenuti mediante assistenza estrattiva ricevono praticamente sempre il massimo dei voti in termini di informatività, qualità, coerenza e attribuibilità.

- La modalità ensemble-GPT-3.5 presenta un rate di compressione medio globale più elevato, pari a circa il 96%, rispetto al full-GPT-3.5, che è circa del 93%. Ciò conferma che i riassunti finali ottenuti, oltre ad essere di ottima qualità, sono approssimativamente del 3% più compatti.

- L’introduzione di nuove tecnologie può migliorare ulteriormente i risultati ottenuti, portando importanti implicazioni pratiche in una vasta gamma di applicazioni.

GRAZIE PER L’ATTENZIONE