ANALISI DELLE METRICHE DI VALUTAZIONE G-EVAL SU DATASET DI DIALOGO: FED E TOPICAL CHAT

Salvatore Sirica

01

PROBLEMA

I sistemi di Natural Language Generation (NLG) generano dialoghi che necessitano di valutazioni per migliorarne la qualità.

01

02

Le metriche tradizionali (es. BLEU, ROUGE) presentano limiti, non correlando sempre bene con i giudizi umani.

02

OBIETTIVO

Sviluppare un framework di valutazione basato su G-EVAL per misurare la qualità dei dialoghi in modo più affidabile

03

G-EVAL

01 Valutazione centrata sull'uomo: Utilizza GPT-4 con il ragionamento Chain-of-Thought (CoT) per ottenere valutazioni strettamente allineate ai giudizi umani, applicabili a vari compiti NLG, come la generazione di dialoghi e la sintesi testuale.

02 Paradigma di compilazione Form-Filling: Impiega prompt strutturati e passi intermedi CoT per standardizzare i punteggi e fornire un processo di valutazione dettagliato e spiegabile.

03 Punteggi dettagliati e precisi: Introduce una normalizzazione basata sulle probabilità per generare punteggi continui e accurati, superando metriche tradizionali come BLEU e ROUGE in termini di correlazione con i giudizi umani.

04

DATASET UTILIZZATI

FED
(Feedback Evaluation Dataset)

Dataset progettato per valutare la qualità complessiva dei dialoghi. Contiene annotazioni umane basate su metriche come coerenza, pertinenza e naturalezza, utili per analisi di qualità globale.

TC_USR
(Topical-Chat User Study)

Benchmark focalizzato su dialoghi contestuali e argomentati. Misura aspetti specifici come il coinvolgimento (engagingness) e la groundedness, fornendo un contesto ricco e diversificato per le valutazioni.

05

FLUSSO G-EVAL

Fase 2

Generazione dei prompt per valutare le metriche definite

Fase 1

Caricamento del dataset

Fase 3

Invio a G-EVAL per ottenere punteggi

06

FASE 1

(CARICAMENTO DEL DATASET)

I dataset utilizzati FED e TC_USR vengono caricati in formato JSON.
Ogni record include:

- Contesto del dialogo.
- Risposta generata dal sistema.
- Annotazioni umane (Overall Score).

Questi dati costituiscono la base per la generazione dei prompt e l’analisi successiva.

07

FASE 2

(GENERAZIONE DEI PROMPT)

Per ogni dialogo o risposta, viene generato un prompt personalizzato. I prompt definiscono le metriche da valutare (Naturalness, Coherence, Engagingness, Groundedness).

Esempio di sostituzione nei template:

- {{context}} → Dialogo completo.
- {{response}} → Risposta generata.

08

FASE 3
(VALUTAZIONE)

I modelli prompt vengono inviati al modello GPT4. Il sistema analizza il contesto e la risposta per valutare le metriche definite:

- Naturalness, Coherence, Engagingness, Groundedness.

I risultati vengono restituiti sotto forma di punteggi dettagliati e spiegabili. Questo processo automatizzato consente una valutazione rapida e standardizzata dei dialoghi.

09

ESEMPIO DI PROMPT

You will be given a dialogue context, a system response, and a related fact.

Your task is to evaluate the response using the following criteria: Naturalness, Coherence, Engagingness, and Groundedness.

Dialogue Context: {{context}}
Fact: {{fact}}
Response: {{response}}

ESEMPIO DI COT

Evaluation Steps:

1. Carefully read the dialogue context to understand its tone, key points, and overall structure.
2. Review the provided fact and use it to verify the accuracy of the response.
3. Analyze the response to determine if it aligns with the dialogue context and maintains engagement based on the evaluation criteria.
4. Assign a score for each criterion on a scale of 1 to 5, where:
   - `1` indicates the lowest quality,
   - `5` indicates the highest quality.

FLUSSO DI ANALISI

Fase 1

Calcolo delle distanze

Fase 2

Generazione risultati

FASE 1

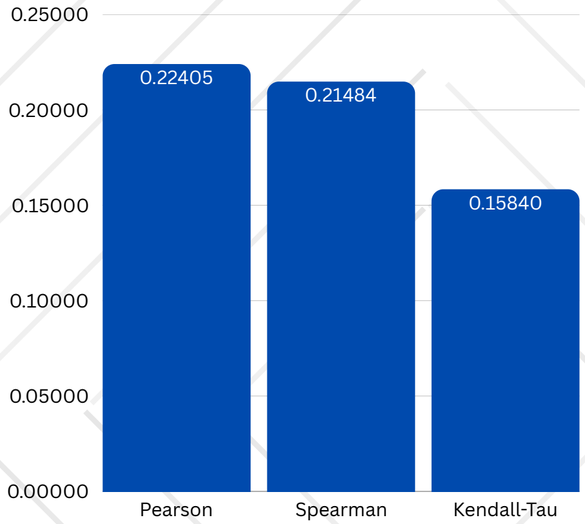
(CALCOLO DELLE DISTANZE)

Valutare la relazione tra i punteggi generati automaticamente (evaluation mean) e quelli forniti dagli annotatori umani (overall score).

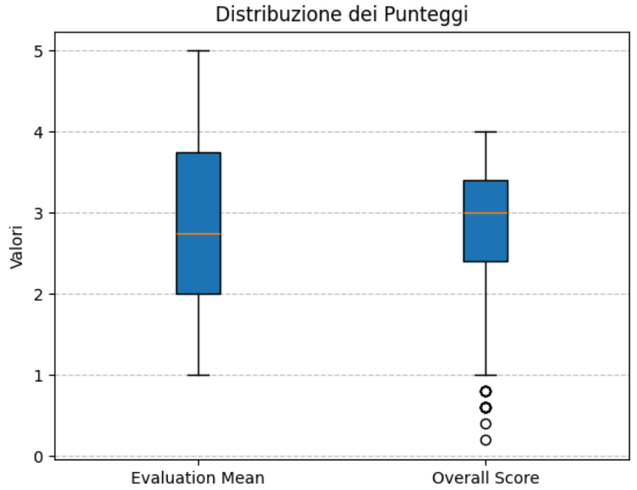
Metriche Utilizzate:

- Pearson: Correlazione lineare tra i punteggi.
- Spearman: Concordanza nei ranghi.
- Kendall-Tau: Monotonicità tra i dati.

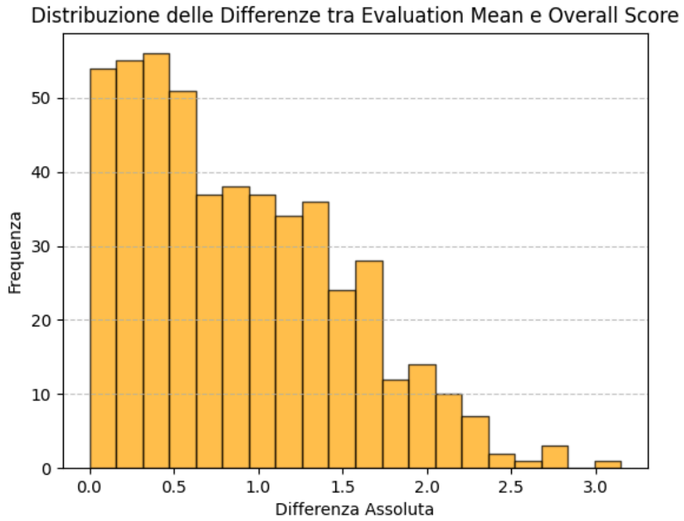
RISULTATI FED



1. Correlazioni Deboli:
Le correlazioni calcolate tra evaluation mean e overall score mostrano valori bassi, suggerendo che le metriche adottate non catturano pienamente i giudizi umani.

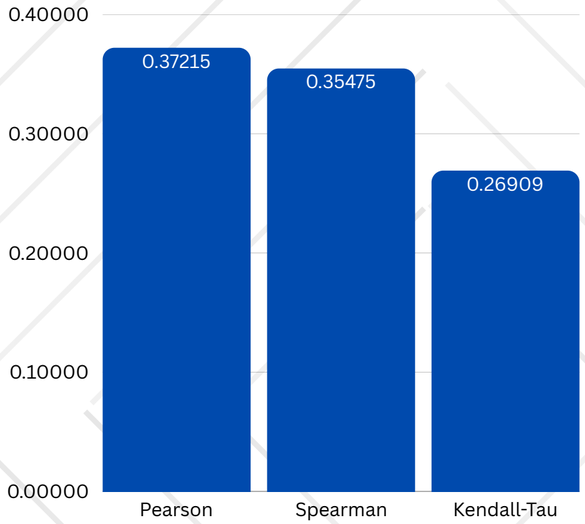


2. Distribuzione dei Valori:
I boxplot mostrano una distribuzione ampia dei punteggi, suggerendo variabilità nel modo in cui le risposte sono state valutate dal sistema rispetto agli annotatori.

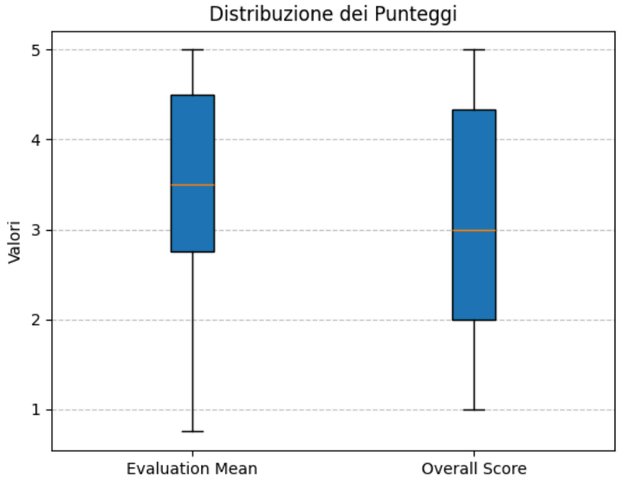


3. Differenze nei Punteggi:
L'istogramma delle differenze evidenzia discrepanze significative tra i punteggi automatici e quelli umani, indicando che il sistema necessita di miglioramenti.

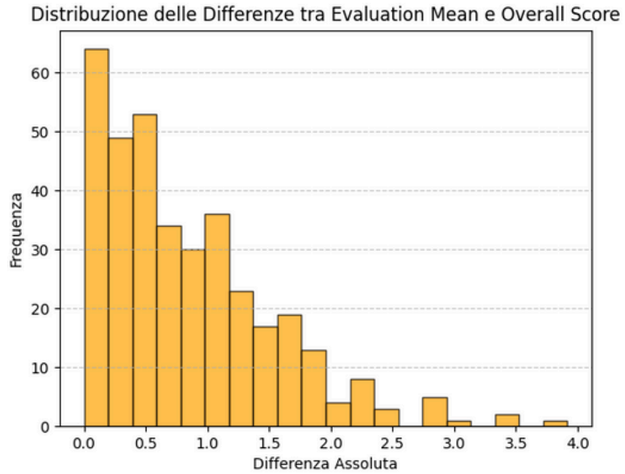
RISULTATI TC_USR



1. Correlazioni Moderate:
Le correlazioni tra evaluation mean e overall score mostrano un miglioramento rispetto a FED, con valori più alti per tutte le metriche.



2. Distribuzione dei Valori:
Il boxplot evidenzia una distribuzione più uniforme rispetto a FED, suggerendo una maggiore coerenza tra le valutazioni automatiche e quelle umane.



3. Differenze nei Punteggi:
L'istogramma delle differenze mostra discrepanze ridotte tra i punteggi generati e quelli annotati manualmente, indicando una valutazione più accurata.

CONCLUSION

01

Limiti Identificati:

- Le discrepanze tra i punteggi automatici e quelli umani suggeriscono margini di miglioramento nelle metriche adottate.
- Le correlazioni, seppur positive, restano deboli in alcuni contesti.

02

Prospettive Future:

- Integrare nuove metriche o modelli basati su reti neurali per aumentare l'affidabilità delle valutazioni.
- Sperimentare ulteriormente su dataset diversificati per validare le metriche su ampi contesti applicativi.

GRAZIE PER L'ATTENZIONE