Analisi dell'Errore

Sebastiano Caccaro

4 ottobre 2021

Sommario

In questo documento è presentata l'analisi dell'errore di sostituzione e la sua correggibilità con sistema di correzione basato sul mask-filling tramite modello BERT.

1 Analisi prestazioni BERT

1.1 Introduzione

Il sistema di correzione sviluppato sfrutta la funzione di mask-filling dei modelli BERT.
In breve, una volta identificato un token non corretto, esso viene mascherato e il sistema propone una serie di proposte, ciascuna accompagnata dalla probabilità che la parola suggerita sia quella corretta. Ad esempio, presa la frase:

"celebrare la venuta dello Spiito Santo, noi vi invitiamo ad implorare da Lui il dono della gioia."

si può notare come il token sottolineato sia un errore. Esso viene quindi mascherato per ottenere la seguente frase:

"celebrare la venuta dello [MASK] Santo, noi vi invitiamo ad implorare da Lui il dono della gioia."

Il sistema propone quindi n candidati per questa la frase data in input, di seguito i primi 5:

- *"Spirito"* con probabilità 0.999
- *"spirito"* con probabilità 7.78e-05

- *"Spazio"* con probabilità 7.36e-06
- *"stesso"* con probabilità 4.31e-06
- *"Zo"* con probabilità 3.85e-06

Per correggere l'errore sono quindi necessarie due condizioni:

- Fra le parole fornite da BERT deve essere presente quella corretta
- Se si verifica la prima condizione, è necessario scegliere la parola corretta

Dalla prima condizione si evince come le prestazioni di BERT siano un limite superiore per ogni sistema che ne usi l'output per effettuare correzioni. Trovando il limite superiore è quindi possibile analizzare in modo più accurato i sistemi di correzione implementati.

1.2 Metodologia dell'analisi

L'analisi dell'errore riguarda solo ed esclusivamente l'ambito della correzione. È quindi necessario che i risultati finali non siano sporcati da altri fattori, come l'error detection.
A questo scopo è stato derivato un nuovo dataset a partire da quello usato per i test di correzione. Innanzitutto vengono estratte delle frasi a campione:

- 3500 per ciascun livello di perturbazione (T1, T2, T3, S1, S2, S3, M1, M2, M3)
- 3500 non perturbate

per un totale di 35000 frasi.
In ognuna di queste frasi viene perturbato un solo token, e viene memorizzata la parola originale. Il token perturbato deve per forza essere presente all'interno di un vocabolario predeterminato, in modo da evitare che token già perturbati vengano ri-perturbati. Il token perturbato viene quindi memorizzato, così come la frase già mascherata. È inoltre memorizzato anche il livello di perturbazione della frase originale.
Viene presentato in seguito un esempio del processo appena descritto. La frase presa a campione

"rarissima', di cui. il Signore si serve per venire a nostro contatto"

viene mappata nel seguente set:

- Frase: *"rarissima', di cui. il Signore [MASK] serve per venire a nostro contatto"*
- Token perturbato: ss
- Token originale: si
- Livello di perturbazione: M2

In questo modo sarà in seguito possibile riprodurre il processo di correzione e confrontare i risultati con la soluzione corretta.
Una volta ottenuto questo dataset, per ogni frase vengono trovati i candidati di correzione proposti da BERT, che vengono aggiunti aggiunti al precedente set insieme alla loro probabilità associata.

1.3 Metriche

L'analisi è mirata a fornire le seguenti statistiche:

- **Correzioni possibili:** percentuale delle volte in cui la correzione corretta è presente nei primi n risultati forniti da BERT. In questo modo è possibile trovare il limite superiore per ogni sistema di correzione che ne faccia uso.
- **Distribuzione delle correzioni:** il ranking di ogni correzione corretta fra i risultati forniti da bert ordinati per probabilità.

1.4 Risultati

Per la misurazione delle correzioni possibili è stato scelto di misurare 3 diverse soglie: viene rilevata la presenza del correzione corretta nei primi 10, primi 20 e primi 30 risultati forniti da BERT. Nei risultati *"text"* rappresenta le frasi non perturbate, mentre *"combinato"* rappresenta le performance sull'intero dataset.
I risultati sono riportati in Figura 1 e Tabella 1.

Perturbazione	Entro 10	Entro 20	Entro 30
text	70.23	74.98	77.41
T1	56.65	62.87	65.42
T2	54.76	59.98	62.47
T3	45.78	51.75	55.03
S1	65.88	71.46	73.77
S2	64.98	69.9	72.24
S3	61.96	67.59	70.62
M1	54.33	59.82	62.94
M2	50.62	56.63	59.76
M3	39.37	45.52	48.51
Combinato	56.47	62.07	64.83

Tabella 1: Correzioni presenti nei primi n risultati prodotti da BERT in percentuale

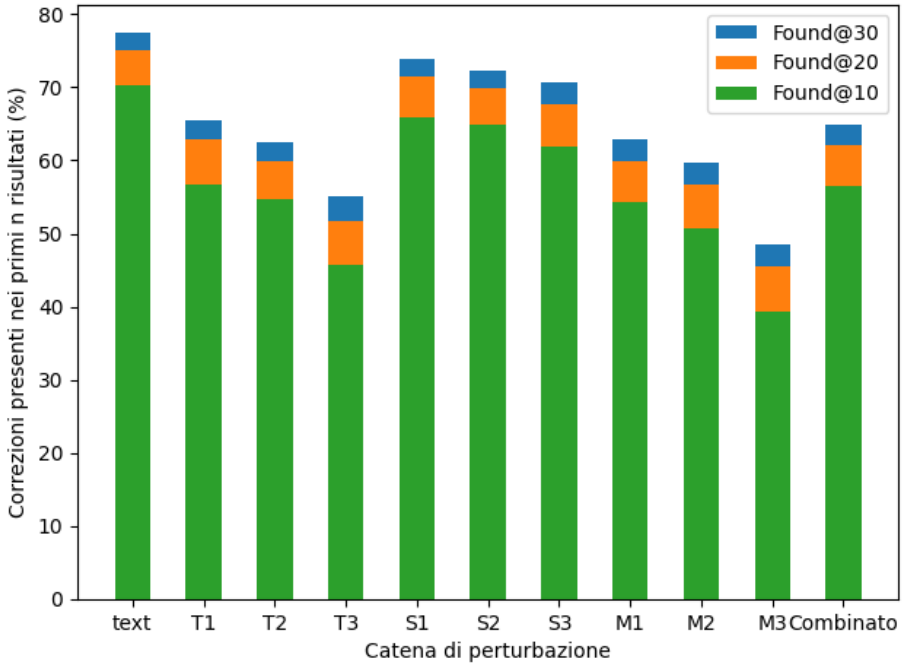

Figura 1: Correzioni presenti nei primi n risultati prodotti da BERT in percentuale

Dai risultati è possibile trarre alcune conclusioni:

- BERT ottiene migliori performance nei testi ottenuti da pipeline con un più alto livello di perturbazione. Questo risultato è lecito da aspettarsi, in quanto le parole perturbate rendono meno intellegibile il contesto che BERT usa per produrre i candidati.
- Mentre l'incremento delle correzioni corrette presenti nei candidati prodotti da BERT è sensibile da 10 a 20 risultati, è praticamente trascurabile da 20 a 30 risultati.

Avendo stabilito la frequenza con la quale la giusta correzione è presente fra i candidati, è necessario che un sistema di correzione sia in grado di trovarla fra la varie alternative proposte.
A tale scopo è utile capire in quale posizione fra i risultati il token corretto si trovi. La distribuzione del posizionamento della soluzione è mostrata nel grafico in Figura 2.

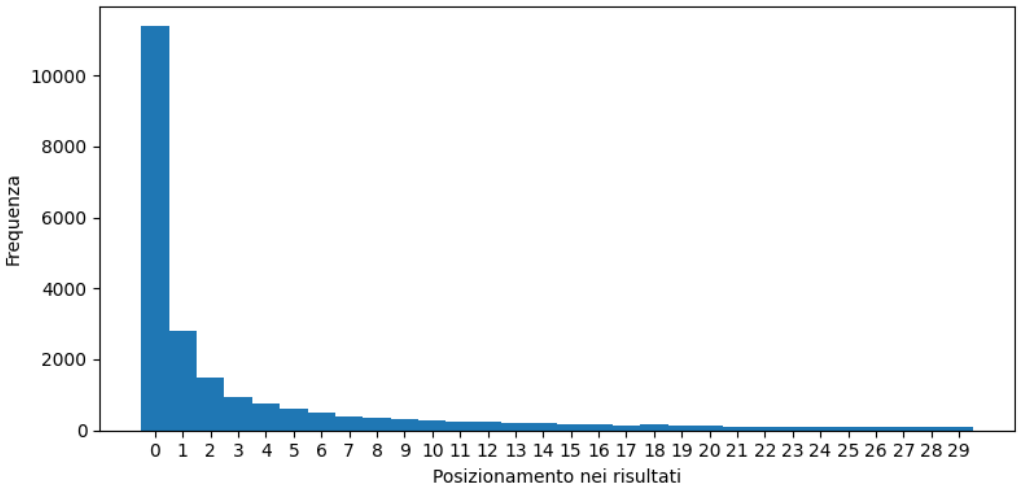

Figura 2: Distribuzione della soluzione nei primi 30 risultati

Il grafico in Figura 2 è relativo al insieme combinato di tutte le frasi. Non vengono riportati i singoli grafici delle altre catene i perturbazione per brevità, in quanto sono praticamente sovrapponili a quello presente.
La distribuzione mostra come i risultati rilevanti si trovino a ridosso

delle prime posizioni. Tuttavia, questa statista non basta per scegliere in modo affidabile il candidato per la correzione.

2 Analisi prestazioni correttore

In questa sezione sono misurate le prestazioni della parte di error correction relativa alla correzione dei singoli token del correttore sviluppato.

2.1 Metodologia dell'analisi

Per testare esclusivamente la parte di error correction viene usata una versione leggermente modificata del sistema di correzione sviluppato:

- La parte di error detection è completamente rimossa, in quanto non necessaria. Infatti, come visto in precedenza, in ogni frase l'errore da correggere è stato introdotto appositamente ed è già demarcato.
- L'output del sistema non è l'intera frase con la correzione incorporata, ma solo la correzione proposta per il token errato.

Una correzione è considerata corretta solo se è esattamente identica alla parola originale.

2.2 Metriche

Per misurare quanto il sistema di correzione riesca ad approssimare il limite superiore di BERT sono introdotte le seguenti metriche:

- **Percentuale di scelta corretta:** per le correzioni nelle quali la soluzione è presente nei primi 20 risultati di BERT, è la percentuale di volte nelle quali il sistema è in grado di identificarla e sceglierla.
- **Percentuale di falsi negativi:** per le correzioni nelle quali la soluzione è presente nei primi 20 risultati di BERT, è la percentuale di volte nelle quali il sistema trova la soluzione, ma la scarta perchè valutata troppo distante dalla parola da correggere.
- **Percentuale di falsi positivi:** per le correzioni nelle quali la soluzione **non** è presente nei primi 20 risultati di BERT, è la percentuale di volte nelle quali il sistema sceglie comunque uno dei candidati, producendo un errore.

2.3 Risultati

Nella Tabella 2 sono riportati i risultati per le metriche definite in precedenza.

Perturbazione	**Scelte Corrette**	**Falsi Negativi**	**Falsi Positivi**
text	88.16	3.05	27.28
T1	86.41	3.59	32.29
T2	83.97	4.01	32.2
T3	80.17	3.14	35.16
S1	86.93	3.93	27.25
S2	87.85	2.84	28.04
S3	86.41	4.06	29.86
M1	85.03	3.77	31.63
M2	84.07	3.88	32.53
M3	80.29	3.58	40.35
Combinato	85.3	3.58	32.5

Tabella 2: Risultati sperimentali dell'error correction del sistema sviluppato

Come è lecito aspettarsi, i risultati sono tanto migliori quanto più basso è il livello di perturbazione.
Come visibile in Figura 3, oltre ai falsi negativi, il sistema sbaglia la soluzione nell'11% delle correzioni. L'eventualità di una correzione sbagliata si presenta quando la soluzione non è la parola più vicina a quella da correggere nei risultati proposti da BERT. Ciò si può evincere anche dal grafico in Figura 4, dove si può notare come la grande maggioranza delle soluzioni sono il risultato più vicino alla parola da correggere.

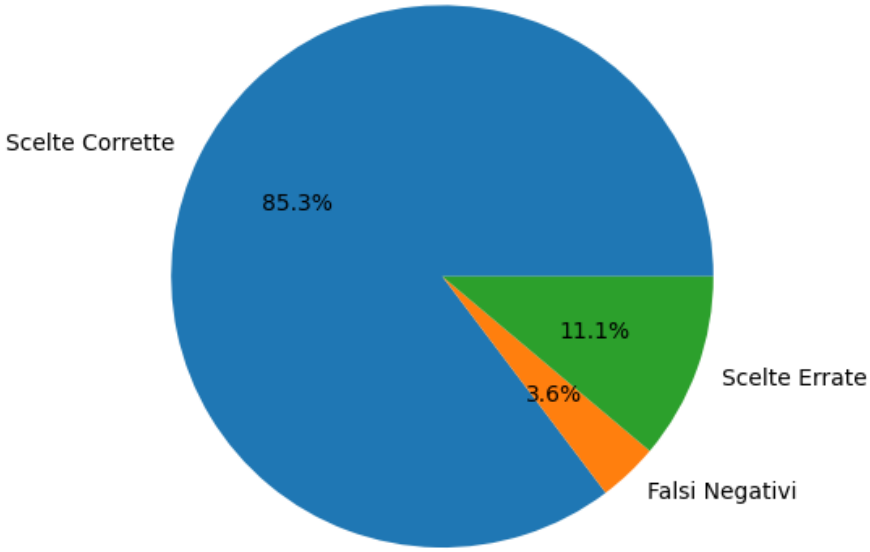

Figura 3: Scelte correte, scelte errate e falsi negativi relativi al combinato di tutti i livelli di perturbazione

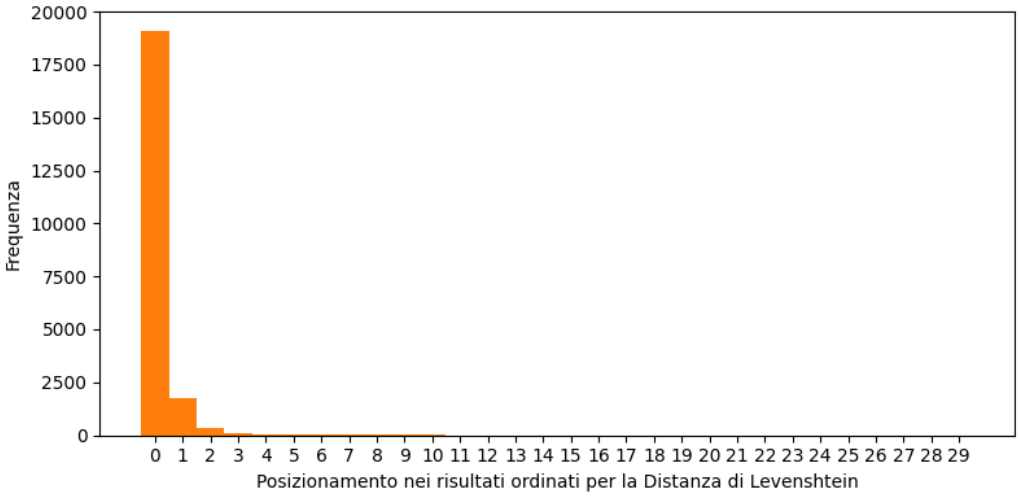

Figura 4: Distribuzione dalla soluzione fra i risultati di BERT ordinati per la distanza di Levenshtein

In Figura 5 invece si può vedere cosa succede in caso la soluzione non sia presente fra i primi 20 risultati: nel 67,5% dei casi il sistema non effettua una correzione, mentre nel resto dei casi viene fornita una correzione errata.

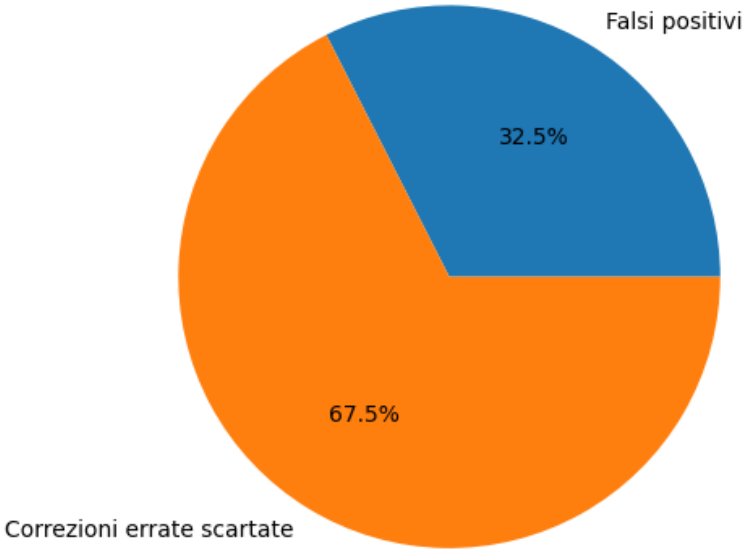

Figura 5: Falsi positivi e correzioni errate scartate relativi al combinato di tutti i livelli di perturbazione

Date queste statistiche, è possibile valutare in modo comprensivo il comportamento del correttore.
Vengono divise le scelte del correttore in due categorie:

- **Scelte positive**: distinguono i casi in cui il correttore sceglie la giusta soluzione se presente, o decide di non intervenire nel caso la soluzione non sia presente fra opzioni fornite da BERT.
- **Scelte Negative**: distinguono i casi in cui il correttore sceglie una correzione sbagliata o decide di non intervenire in caso la soluzione sia presente fra i risultati di BERT. Sono inclusi anche i casi nei quali, in mancanza della soluzione fra i risultati, il correttore fornisce una correzione errata.

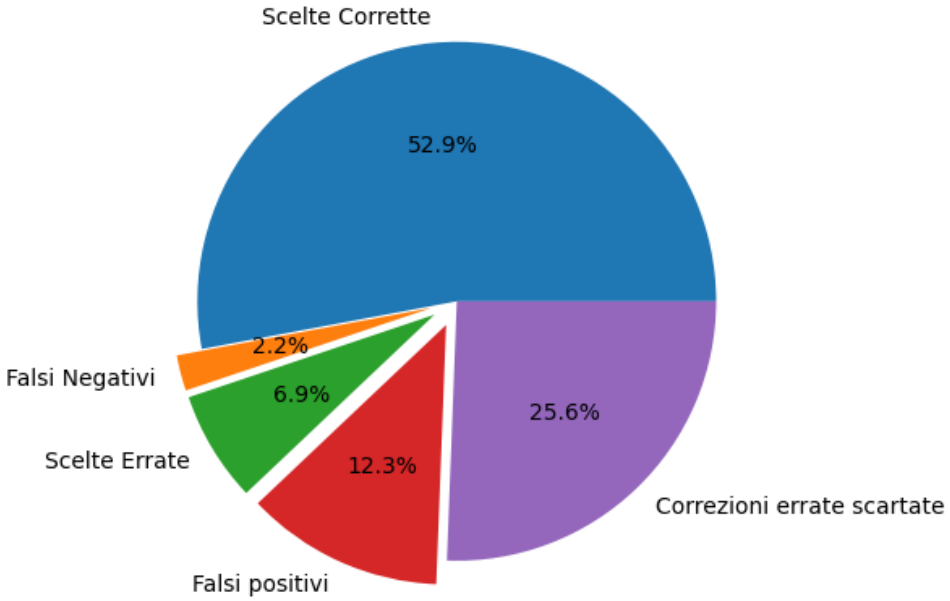

Figura 6: Scelte positive e negative relative al combinato di tutti i livelli di perturbazione. Nel diagramma, le scelte negative sono quelle distaccate dalla torta principale.

Dal diagramma in Figura 6 si nota come il sistema di correzione faccia un scelta positiva nel 78% dei casi.
In Tabella 3 sono invece riportate le percentuali di scelte negative e positive per tutti i livelli di perturbazione.

Perturbazione	**% Scelte Positive**	**% Scelte Negative**
text	84.29	15.71
T1	79.47	20.53
T2	77.5	22.5
T3	72.77	27.23
S1	82.88	17.12
S2	83.06	16.94
S3	81.14	18.86
M1	78.34	21.66
M2	76.87	23.13
M3	69.04	30.96
Combinato	78.55	21.45

Tabella 3: Percentuale di scelte positive e negative per tutti i livelli di perturbazione