Analisi dell'Errore

Sebastiano Caccaro

1 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

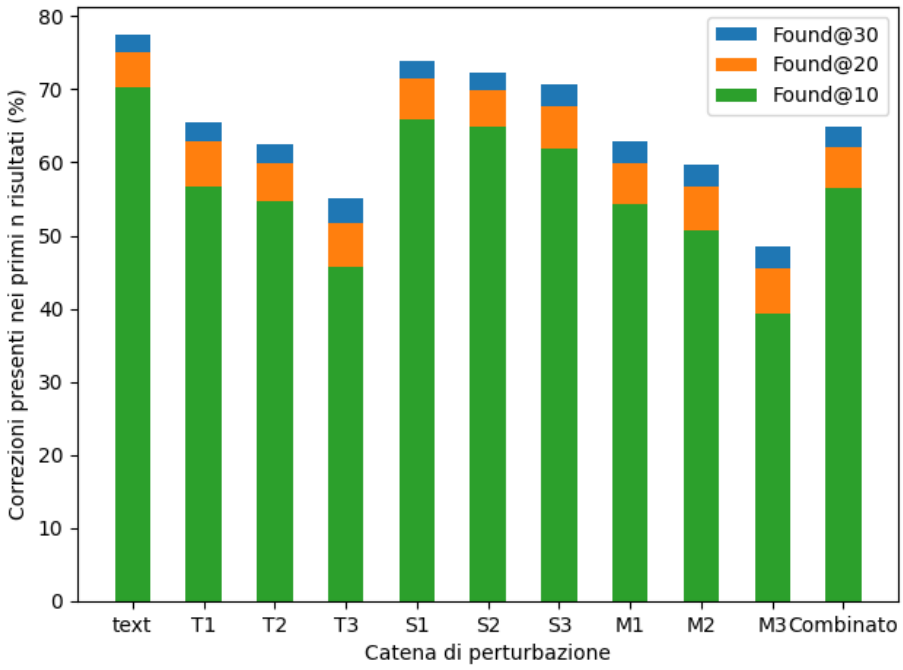

Figura 1: Correzioni presenti nei primi n risultati prodotti da BERT in percentuale