Perturbazione di testo

Sebastiano Caccaro

7 giugno 2021

1 Introduzione

In questo documento è descritto il funzionamento del perturbatore di testo. Il perturbatore è un funzione che, dato un un determinato testo come input, ne produce una versione perturbata tramite l'introduzione alcuni errori.
Tali errori sono modellati attraverso delle funzioni, chiamate moduli, che vengono composte tra loro per creare una funzione detta pipeline. Più formalmente:

$$Pipeline : m_k \circ m_{k-1} \circ m_{k-2} \circ ... \circ m_0 \tag{1}$$

dove $\{m_0, ..., m_k\}$ sono moduli.
Il perturbatore può essere visto come una combinazione di n pipeline, dove ogni pipeline P_i è associata ad un peso w_i. Il testo in input x viene diviso in una sequenza di blocchi $\{x_1, ..., x_k\}$ dove per ogni blocco viene applicata una sola pipeline P_i con una probabilità

$$p_{prt_i} = \frac{w_i}{\sum_{j=0}^{n} w_j} \tag{2}$$

Inoltre, viene definito il parametro del perturbatore $s \in \{0...1\}$, detto stickyness. Tale parametro comporta che, dato un blocco x_h a cui è applicata la pipeline P_i, al blocco x_{h+1} sia applicata la stessa pipeline con probabilità:

$$p_{stick_i} = s + (1 - s)p_{prt_i} \tag{3}$$

2 Moduli

La perturbazione del testo avviene attraverso diverse unità chiamate moduli. Sono definiti tre tipi di moduli:

- Moduli di tokenizzazione
- Moduli di detokenizzazione
- Moduli di perturbazione

2.1 Moduli di tokenizzazione

I moduli di tokenizzazione sono delle funzioni che hanno lo scopo di scomporre del testo in diversi token. Più formalmente un modulo di tokenizzazione è definito come segue:

$$T : x \mapsto \{t_1, t_2, ..., t_n\} \tag{4}$$

dove t_i è un token.
Esempio
Input: Nel mezzo del cammin di nostra vita mi ritrovai per una selva oscura, ché la diritta via era smarrita.
Output: {'Nel', 'mezzo', 'del', 'cammin', 'di', 'nostra', 'vita', 'mi', 'ritrovai', 'per', 'una', 'selva', 'oscura', ',', 'ché', 'la', 'diritta', 'via', 'era', 'smarrita', '.'}

2.2 Moduli di detokenizzazione

I moduli di detokenizzazone sono delle funzioni che hanno lo scopo di mettere insieme una lista ordinata di token in un'unica stringa di testo. Sono definiti come segue:

$$D : \{t_1, t_2, ..., t_n\} \mapsto x \tag{5}$$

Esempio
Input: {'Nel', 'mezzo', 'del', 'cammin', 'di', 'nostra', 'vita', 'mi', 'ritrovai', 'per', 'una', 'selva', 'oscura', ',', 'ché', 'la', 'diritta', 'via', 'era', 'smarrita', '.'}
Output: Nel mezzo del cammin di nostra vita mi ritrovai per una selva oscura, ché la diritta via era smarrita.

2.3 Moduli di perturbazione

I moduli di perturbazione sono delle funzioni che hanno lo scopo di inserire gli errori all'interno del testo. Un modulo di perturbazione è definito come segue:

$$P : \{t_1, t_2, ..., t_n\} \mapsto \{t'_1, t'_2, ..., t'_k\} \tag{6}$$

Ogni modulo P è caratterizzato dai seguenti parametri:

- $f : \{t_1, t_2, ..., t_g\} \mapsto \{t'_1, t'_2, ..., t'_h\}$. Una funzione che prende in input gruppi di g token e restituisce h token perturbati.
- p: probabilità che un dato gruppo di token $\{t_1, t_2, ..., t_g\}$ sia perturbato dalla funzione f.

A seconda della funzione f, si possono distinguere varie categorie di moduli di perturbazione, che modellano diversi errori. Inoltre, alcuni moduli possono richiedere alcuni parametri aggiuntivi.

2.3.1 Modulo di split

Il modulo di split modella l'errore in cui le lettere di una parola vengono intermezzate da degli spazi. La funzione è definita come:

$$Split : \{t_1\} \mapsto \{t'_1\} \tag{7}$$

Essendo t_1 un token, e quindi una sequenza di caratteri, definiamo t'_1 come:

$$t'_1 = c +' \ ' \forall c \in t_1 \tag{8}$$

Esempio
Input: {`'cammin'`}
Output: {`'c a m m i n'`}

2.3.2 Modulo di aggiunta di punteggiatura

Il modulo di aggiunta di punteggiatura modella l'aggiunta di un segno di punteggiatura come il punto, la virgola o un apostrofo fra due token. La funzione è definita come:

$$AddPunct : \{t_1\} \mapsto \{t_1, punct\} \tag{9}$$

dove *punct* è parametro della funzione *AddPunct*.
Esempio
Input: {`'cammin'`}
Output: {`'cammin',','`}

2.3.3 Modulo di unione con trattino

Il modulo di unione con trattino modella l'unione di due token attraverso l'uso del carattere '-'. È definito come segue:

$$MergeHypen : \{t_1, t_2\} \mapsto \{t_1 + \text{'-'} + t_2\} \tag{10}$$

Esempio
Input: {'del','cammin'}
Output: {'del-cammin'}

2.3.4 Modulo di divisione con virgola

Il modulo di divisione con virgola modella l'inserimento di una o più virgole all'interno di un token. È definito come segue:

$$SplitComma : \{t_1\} \mapsto \{f(t_1)\} \tag{11}$$

Definendo un token t come una sequenza di caratteri $\{c_1, ..., c_n\}$, e definendo un numero casuale $x \in \mathbb{N} \wedge 1 \leqslant x < n$, è possibile formalizzare f come:

$$f(t) = \begin{cases} \{c1, ..., c_x\} + \text{','} + f(\{c_{x+1}, ..., c_n\}) & \text{se } x < n \\ '' & \text{se } x \geqslant n \end{cases} \tag{12}$$

Esempio
Input: {'cammin'}
Output: {'ca,mm,in'}

2.3.5 Modulo di rimpiazzo caratteri

Il modulo di rimpiazzo caratteri modella lo scambio di una sequenza di caratteri con un'altra all'interno dello stesso token. È definito come segue:

$$SubChar : \{t_1\} \mapsto \{f(t_1)\} \tag{13}$$

La funzione f è caratterizzata dai seguenti parametri:

- Un insieme di sequenze di caratteri $\{s_1, ..., s_n\}$ dove ad ogni sequenza s_i è associata una probabilità p_i. Ogni sequenza s_i presente in un token t viene rimpiazzata da un'altra sottosequenza con probabilità p_i.

- Ad ogni sequenza s_i è associato un insieme di sequenze $\{r_{i1}, ..., r_{ik}\}$ dove ogni sequenza s_{ij} è associata ad un peso w_{ij}.

Se ad ogni probabilità p_i viene associata una variabile $X_i \sim Ber(p_i)$, e definisco S come l'insieme di tutte le sequenze s_i all'interno di t, è possibile definire f come:

$$f(t,S) = \begin{cases} t & \text{se } S = \emptyset \\ f(t, S \setminus \{s_i\}) & \text{se } S \neq \emptyset \wedge X_i = 0 \\ f(\{c_1, ..., c_g\}, S) + sub(\{c_{g+1}, ..., c_h\}) + f(\{c_{h+1}, ..., c_n\}, S) & \text{se } S \neq \emptyset \wedge X_i = 1 \end{cases} \tag{14}$$

dove se $\{c_{g+1}, ..., c_h\}$ è la sequenza s_i, $sub(\{c_{g+1}, ..., c_h\})$ è una sola fra le sequenze $\{r_{i1}, ..., r_{ik}\}$ scelta con probabilità

$$p_{sub_{ij}} = \frac{w_{ij}}{\sum_{h=0}^{k} w_{ij}} \tag{15}$$

Esempio
Input: {'cammin'}
Output: {'oanimin'}