

Valutazione test

Sebastiano Caccaro

4 luglio 2021

Sommario

In questo documento esposti i risultati dei test ottenuti con diversi livelli e tecniche di perturbazione. Sono inoltre usati diversi metodi di correzione o più versioni dello stesso metodo.

1 Metodi

In questa sezione sono descritte le configurazioni di ogni test eseguito in termini di dataset e di metodo di correzione. Ogni codice è inoltre caratterizzato da un identificativo univoco. Ogni identificativo è composto come segue:

xxxyyz

dove:

- xxx è una stringa di tre lettere che individua il metodo di correzione utilizzato.
- yy è un numero che indica una determinata configurazione dei parametri del metodo utilizzato.
- z è una lettera che indica un diverso metodo di perturbazione utilizzato nel dataset.

2 Definizioni configurazioni

2.1 Metodi di correzione

I metodi di correzione usati sono i seguenti:

- **Project-gender-politics (pgp)**: metodo sviluppato da Francesco Periti. Nome lasciato così per mancanza di alternative. Sono presenti le seguenti configurazioni:
 - 00: ottenuta allenando i modelli word2vec e gensim con i parametri di default.
 - 01: ottenuta allenando i modelli word2vec e gensim con i seguenti parametri:
 - * min_count = 5
 - * window = 5
 - * vector_size = 100
 - * sample = 1e-4
 - * negative = 20
 - * epochs = 50
 - * min_alpha = 0.0001
 - * alpha = 0.025

2.2 Dataset

I tipi di dataset utilizzati sono i seguenti:

- **a**: gli errori di tokenizzazione sono introdotti attraverso il modulo *CharSub*.
- **b**: gli errori di tokenizzazione sono introdotti attraverso il modulo *TokenSub*.

3 Metriche

4 Risultati

I risultati per la metrica x... sono.

	T1	T2	T3	S1	S2	S3	M1	M2	M3
pgp00a	0.139	0.150	0.160	0.108	0.122	0.132	0.145	0.154	0.153