

Abstract

“La comunicazione parlata per addestrare i Language Models. Sfide e opportunità”
di Alessandro Lenci

L'implausibilità cognitiva dei Large Language Models (LM) è un argomento ricorrente nella letteratura sull'elaborazione del linguaggio naturale (NLP). Da un punto di vista puramente quantitativo, i Large LM sono addestrati su corpora che contengono diversi ordini di grandezza di parole in più rispetto all'input linguistico a cui è esposto un essere umano. Da un punto di vista qualitativo, i dati di addestramento dei Large LM sono costituiti principalmente da testi scritti derivati dal web. Questo articolo presenta BAMBI (BABy language Models Boostrapped for Italian), una serie di Baby Language Models (BabyLMs) addestrati su dati che imitano l'input linguistico ricevuto da bambini in età prescolare o elementare e che consistono principalmente in dati di comunicazione parlata trascritti. Verranno discusse le sfide poste dalla raccolta di questo tipo di dati e le opportunità che essi possono offrire per sviluppare LM più plausibili dal punto di vista cognitivo da utilizzare per studiare l'acquisizione e l'elaborazione del linguaggio.