IUSS Scuola Universitaria Superiore Pavia
Sezione:IUSS PAVIA

IL FOGLIO
Dir. Resp.:Claudio Cerasa
Tiratura: 25.000 Diffusione: 25.000 Lettori: 120.000

Rassegna del: 28/04/21
Edizione del:28/04/21
Estratto da pag.:7
Foglio:1/1

E se i dati fossero "too big" per capire il linguaggio? Dubbi in Google

Tiro in alto una moneta: viene testa. Concludo: tutte le monete tirate in alto ricadono sulla testa. Nemmeno un ignorante di statistica come me abboccherebbe. Per avere una valutazione veritiera devo tirare in alto la moneta e farla ricadere tante e tante e tante volte (per poi scoprire che solo la metà delle volte in media cade sulla testa, altrimenti non direi che una cosa imprevedibile è come tirare la monetina). A che punto devo fermarmi nella raccolta dei dati? Con la questione della monetina non dovrebbe essere difficilissimo saperlo ma, se la regolarità che voglio catturare su base statistica è più complicata di una caduta su un lato o un altro di una monetina, la risposta non è così evidente.

Intanto, devo prendere una decisione esplicita e decidere se una certa regolarità, un fenomeno, detto in termini più trasparenti, si può cogliere in modo statistico. E qui, immediatamente, iniziano i problemi. E' invalso in molti l'atteggiamento secondo il quale qualsiasi fenomeno, a patto di avere un numero sufficiente di dati, può essere colto in modo statistico. Intendiamoci: forse questo è vero per un Dio Laplaciano che vede il mondo come un ciclopico carillon. Anche la teoria eliocentrica del sistema solare, per chi lo vede così, può forse essere colta scattando di notte miliardi di miliardi di fotografie e poi facendo osservare, si fa per dire, le fotografie a una macchina che possa trovare la correlazione minima tra i puntini luminosi che spieghi i dati. Forse: e non è detto né che bastino le risorse di computo né che basti il tempo per aspettare. Comunque sia, le cose non sono andate storicamente così e la teoria eliocentrica non è stata scoperta su base statistica esattamente come la legge di gravità di Newton o la teoria della trasmissione genetica di Mendel. Quando si parla di linguaggio, invece, questo miraggio della comprensione automatica continua a rigenerarsi. Si palesò la prima volta negli anni Cinquanta, quando si aprirono le porte dei laboratori bellici con le macchine per la decifrazione dei codici segreti come il famosissimo Enigma, utilizzato dai nazisti: in quegli anni fiorì l'illusione della traduzione automatica e con essa della comprensione delle grammatiche su base statistica. Poi i lavori di matematica del giovane Noam Chomsky alla fine di quel decennio mitico per la scienza confutarono queste possibilità a suon di teoremi ma il miraggio, come fosse dotato di vita propria al pari di un fantasma, è ritornato, complice una colpevole confusione tra comprensione e simulazione di un comportamento umano e l'enorme accrescimento della capacità di calcolo delle macchine elettroniche.

Morale: negli anni recenti è tornata l'idea che, a patto di avere un numero cospicuo (big, per gli anglofoni lessicopenici) di dati, si possa decifrare la struttura di una lingua umana (e dunque di tutte le lingue umane) semplicemente utilizzando la statistica. Infischiandosene delle prove dei teoremi chomskyani e di ormai una sessantina di anni di ricerca formale e comparate delle lingue umane sulla base del metodo analitico che ne è scaturito (grammatica generativa), si è tornati a pensare che se si hanno dei big data (curioso, l'accostamento del plurale neutro latino all'aggettivo inglese) tutto si dipana e la struttura delle lingue umane è a portata di mano, anzi di macchina. Ma la domanda si ripropone, come per la monetina: quanto big? La risposta non è nemmeno tanto astratta, visto che a seconda della grandezza del repertorio servono soldi a disposizione: per crearlo e per valutarlo in modo automatico.

La storia finirebbe qua se non fosse che una decina di giorni fa Samy Bengio, prestigioso e stimatissimo scienziato informatico, fondatore e condirettore del progetto Google Brain, si è dimesso dopo 14 anni. Questo gesto, clamoroso quanto inaspettato, segue il licenziamento da parte della stessa società di una ricercatrice, Timnit Gebru, per la pubblicazione di un articolo che mette in forse la capacità di comprendere il linguaggio a partire da una messe sufficientemente grande di dati. Evidentemente deve essere stato insostenibile per una big cathedral fondata sull'abbondanza di dati come la Google criticare questa strategia. Il sottotitolo dell'articolo pietra dello scandalo, tuttavia parla chiaro: "E' possibile che i modelli per il linguaggio siano troppo grandi?". Stai a vedere che la vecchia strategia di Chomsky secondo il quale per capire come funziona il linguaggio non occorrono fantastiliardi di dati e di macchine ma basta avere come modello il cervello di un bambino o, forse meglio, di una bambina, e i pochi frammentari stimoli che riceve nei primi tre o quattro anni di vita si riveli ancora una volta il modello vincente. E se il sottotitolo dell'articolo è chiaro, il titolo lo è ancora di più e vale la pena impararselo a memoria: "Sul pericolo dei pappagalli stocastici". Ancora una volta noi esseri umani non risultiamo facilmente replicabili.

Andrea Moro
Scuola universitaria superiore Iuss, Pavia


ME CONTRO TECH

Peso:20%