Giganti della lingua e risorse limitate
di Paolo Caffoni

Pubblicato in
Medial Disorders: Interpretative and Non-statistical Compendium of Technological Disorders
Inactual 2024

1. Non lasciare nessuno indietro

Nel 2022, Meta lancia in opensource No Language Left Behind 200 (NLLB200), un modello di traduzione automatica multilingue in grado di fornire traduzioni direttamente tra qualsiasi coppia di oltre 200 lingue. Fra queste, 146 sono le cosiddette "lingue a bassa disponibilità di risorse" (Low Resource Languages), come ad esempio il ligure (parlato da circa 600.000 persone), l'arabo levantino meridionale (o arabo palestinese) e il wolof (circa 12 milioni di parlanti nell'Africa occidentale) (NLLB, 2022).

Come sono definite le lingue a bassa disponibilità di risorse? Queste sono considerate lingue meno studiate, carenti di risorse, meno informatizzate, meno privilegiate, meno comunemente insegnate o a bassa densità, tra altre definizioni (Cieri et al., 2016; Magueresse et al., 2020). NLLB200 le definisce come le lingue che non raggiungono 1 milione di esempi di frasi tradotte e pubblicamente disponibili online, rendendole così inadatte per l'applicazione diretta di metodi statistici per la traduzione automatica a causa della scarsità di dati rilevabili.

In un'intervista pubblicata online[1], Angela Fan, ricercatrice di Meta, dichiara che l'idea di NLLB200 è stata ispirata al sentimento espresso in molte delle interviste nella fase preliminare del progetto: la mancanza di accesso alle tecnologie di traduzione automatica preclude l'accesso a nuove opportunità, riporta Fan, che continua esemplificando: "Anche se posso essere la persona più intelligente del mio paese, ho comunque accesso a molto meno materiale letterario e di apprendimento digitale rispetto a un mio coetaneo cresciuto in un paese anglofono". Nonostante tutto, la domanda rimane: perché un'azienda da miliardi di dollari di fatturato come Meta è interessata ad automatizzare la traduzione del ligure?

2. Monolinguismo digitale

Seguendo il ragionamento del linguista Nicholas Ostler, potremmo prevedere che sia solo una questione di tempo prima che l'ultima lingua franca (ovvero l'inglese) venga sostituita dall'ubiquità delle tecnologie di traduzione automatica (Ostler, 2010). Per Ostler, se è vero che modelli e corpora linguistici hanno, fino ad oggi, privilegiato le lingue storiche dell'Europa occidentale, non passerà molto tempo prima che anche lingue a cosiddetta "bassa densità" raggiungano gli standard imposti dai colonizzatori. Se questo accadrà a beneficio – alternativamente, e a seconda delle occasioni – dell'influenza esercitata dai pionieri o della sopravvivenza di chi rincorre, è ancora difficile da prevedere. «In definitiva [scrive Ostler] e forse fra non troppo tempo, diciamo verso la metà del ventunesimo secolo, tutti potranno esprimere un'opinione nella propria lingua, sia oralmente che per iscritto, e il mondo capirà» (Ostler, 2010).

1 Angela Fan explains NLLB-200: High-quality Machine Translation for Low-Resource Languages. https://www.youtube.com/watch?v=IJZE7LikM3c.

In ultima analisi, quella di Ostler sembra essere una prognosi meno azzardata se consideriamo che la traduzione automatica e le varie lingue franche (ad esempio il latino, l'inglese o il malese) o lingue artificiali, come l'esperanto, hanno mostrato una comune tendenza a eliminare le barriere linguistiche affrontando il problema della traduzione da un punto di vista monolinguale, cioè quello di dire: non ci sarà una necessità assoluta di imparare lingue "straniere" in futuro, poiché tutti potranno comunicare utilizzando ponti di mediazione, siano questi dati in forma utopica di lingue universali, o tecnologica, come il servizio di traduzione automatica delle telefonate proposto da Samsung[2].

Negli ultimi anni, a seguito dell'ampio successo dei Modelli linguistici a grandi dimensioni, come GPT e BERT, l'inadeguatezza di un approccio monolinguale alla traduzione automatica è stato esteso a preoccupazioni riguardanti le rappresentazioni monoculturali e i pregiudizi (*bias*) incorporati nel Machine Learning. In un saggio ampiamente circolato e discusso del 2021, Emily M. Bender e Timnit Gebru hanno messo in guardia sui rischi connessi all'utilizzo incrementale di queste tecnologie. Fra questi, «la tendenza dei dataset prelevati da internet a codificare visioni del mondo egemoniche, la tendenza dei Modelli linguistici di grandi dimensioni ad amplificare pregiudizi e altri problemi già presenti nei dataset, e la tendenza dei ricercatori a confondere incrementi nelle performance dei modelli statistici come una reale comprensione del linguaggio naturale» (Bender and Gebru et al., 2021).

Ad oggi, Bender e Gebru riportano che solo il 7% dei dati di addestramento per un modello multilingue come GPT-3 è in una lingua diversa dall'inglese, mentre circa il 90% delle lingue del mondo, utilizzate da un miliardo di persone, attualmente ha scarsa o nessuna assistenza quando si tratta di tecnologie linguistiche. In questo senso, la crescita esponenziale dei dataset utilizzati per addestrare i Modelli linguistici di grandi dimensioni non ha corrisposto a una diversificazione delle visioni del mondo rappresentate. Tutt'al più, l'impatto di un aumento dell'accuratezza nella traduzione dall'inglese al tedesco dello 0,1 secondo il punteggio BLEU (una delle metriche più popolari per valutare la traduzione automatica) corrisponde a un aumento di $150.000 nei costi di computazione e delle emissioni di carbonio a esso correlate (Strubel et al., 2019).

3. Preservare la "biodiversità" linguistica

Per comprendere appieno gli sforzi in corso per diversificare i dataset e le prestazioni dei Modelli linguistici di grandi dimensioni, è insufficiente fare affidamento esclusivamente su un discorso critico verso i *bias* incorporati nei modelli statistici. Come è anche ugualmente insufficiente prognosticare un inevitabile tramonto dell'ultima lingua franca verso un ritorno alla diversità multilingue di Babele.

Un ulteriore aspetto che dovrebbe essere preso in considerazione è l'ampia influenza esercitata da rappresentazioni figurate come quelle di "lingue in pericolo di estinzione" o "diversità bioculturale", che apparentemente traggono significato dall'associazione con i topoi di "conservazione" e "biodiversità" nelle scienze naturali. L'equiparazione di diversità biologica e culturale-linguistica, assieme ai loro valori intrinseci, pone una questione sulla legittimità del riferimento metaforico all'estinzione in campo linguistico, e sulla definizione di organico e

2 Vedi: https://www.ilpost.it/2024/04/18/intelligenza-artificiale-lingue-straniere/

organismo in relazione alla lingua. Dove ci porta il parallelo tra "risorsa linguistica" e "risorsa naturale"?

Lo storico della scienza David Sepkoski sostiene che il termine "diversità bioculturale" sia emerso nella seconda metà del ventesimo secolo, a seguito di un rinnovato interesse per il problema dell'estinzione e di un apprezzamento per la diversità delle risorse biologiche e culturali necessarie a garantire la sopravvivenza dell'umanità. «Nella sua formulazione di base [scrive Sepkoski] l' argomento a favore della diversità biologica in quanto 'risorsa' è significativamente cambiato poco per oltre quarant'anni. Che siano intese concretamente come risorse materiali tangibili, prodotti alimentari, medicinali, beni economici o, più astrattamente, come 'informazioni', ad esempio, informazioni genetiche, è stato tendenzialmente il valore 'utilitaristico' della diversità biologica ad attirare l'attenzione» (Sepkoski, 2020). L'accento sulla diversità linguistica ha trovato il suo culmine nella conferenza tenutasi a Berkeley nel 1996 "Endangered Languages, Endangered Knowledge, Endangered Environments", sponsorizzata dall'UNESCO e dalla neo-fondata fondazione Terralingua, durante la quale elementi chiave nella comprensione della diversità biologica hanno influenzato le discussioni sul pericolo di estinzione anche in ambito culturale e linguistico (Maffi, 2001).

È interessante notare che gli sforzi per rispondere all'emergenza portata da una "crisi delle risorse linguistiche" hanno visto l'uso dell'Elaborazione del linguaggio naturale (NLP, da natural language processing) e della traduzione automatica sia come origine che come soluzione al problema in questione.

Il linguista computazionale András Kornai, fra i primi, ha lavorato per estendere la diagnosi culturalista sui rischi dell'estinzione delle lingue anche alla sfera digitale, proponendo la definizione di "morte digitale della lingua". Utilizzando tecniche di machine learning, Kornai ha classificato le lingue su scala globale in base alla loro vitalità digitale, basandosi su criteri come funzione, prestigio e perdita di competenza, concludendo che «delle 7.000 lingue ancora viventi, forse 2.500 sopravviveranno, in senso classico, per un altro secolo, mentre saranno solamente 250 le lingue che sopravviveranno digitalmente» (Kornai, 2013).

Al contrario, Steven Bird e David Chiang hanno promosso l'uso della traduzione automatica a scopo di conservazione nel caso di 15 lingue in pericolo di estinzione nelle regioni montuose della Papua Nuova Guinea, fra le aree più diversificate al mondo a livello linguistico. Per evitare che le lingue cadano in disuso prima che i linguisti possano documentarle, e tentando di aggirare la scarsità di risorse umane ed economiche attualmente a disposizione, Bird e Chiang propongono di fornire ai parlanti indigeni una metodologia di documentazione basate sui modelli statistici di traduzione automatica (Bird e Chiang, 2012).

Questi approcci tecno-deterministici appaiono in netto contrasto con l'attenzione critica che alcuni studiosi indigeni hanno rivolto a definizioni retoriche come quelle dell'"ultimo parlante" e delle "voci che stanno scomparendo" (Nettle e Romaine, 2000), spesso impiegate dai linguisti occidentali allo scopo allertare la popolazione dell'incombente catastrofe. Sottolineando come quello dell'estinzione linguistica sia un "concetto invasivo" di eredità coloniale, gli studiosi delle lingue indigene americane hanno esortato a discutere una nuova configurazione della sovranità linguistica che promuova la prospettiva del "risveglio" e della "vitalità linguistica", in contrasto con gli effetti

destabilizzanti sulle comunità parlanti causati dalle pratiche tassonomiche di documentazione e archiviazione dei codici linguistici (Baldwin, et al., 2018).

4. Taci e lavora? Anzi, parla!

Ancora prima che diversità biologiche e culturali diventassero temi centrali nel dibattito sull'Antropocene, l'idea che le macchine potessero fare ricorso a "risorse linguistiche" aveva trovato risonanza nei primi anni Novanta in alcune analisi del post-operaismo italiano sul ruolo centrale occupato dal linguaggio nella riconfigurazione della divisione del lavoro (Virno, 2001[3]; Marazzi, 1994).

In polemica con la distinzione habermasiana tra "agire strumentale" (Arbeit) e "agire comunicativo" (Interaktion), Paolo Virno osservò nella cooperazione linguistica tra donne e uomini, nel loro concreto "agire di concerto" all'interno di nuovi tipi di fabbriche, come la FIAT a Melfi negli anni Novanta, un'articolazione tra conoscenza e produzione che non poteva essere risolta solo all'interno del capitale fisso. Scriveva Virno: «Nei processi lavorativi contemporanei, vi sono pensieri e discorsi che funzionano di per sé come 'macchine' produttive, senza dover adottare un corpo meccanico e neppure una animella elettronica» (Virno, 2001). Quando i lavoratori sono incaricati di inventare nuove procedure cooperative, l'atto di integrazione linguistica viene messo in primo piano e costituisce la principale risorsa delle appropriazioni capitaliste nella "fabbrica loquace".

Nel descrivere la transizione verso un'economia post-fordista, l'economista Christian Marazzi fa eco all'ipotesi di Virno e argomenta che l'ingresso diretto della comunicazione nel processo produttivo è derivato da un rovesciamento della relazione storica tra le sfere della produzione e quella della distribuzione: nel sistema fordista, la produzione escludeva la comunicazione, poiché la catena di montaggio era essenzialmente muta ed eseguiva meccanicamente le istruzioni dei colletti bianchi. Nel post-fordismo, assistiamo invece a una catena di montaggio "parlante", caratterizzata da connessioni semantiche, dove la comunicazione e le tecnologie impiegate possono essere considerate vere e proprie "macchine linguistiche" volte a razionalizzare e accelerare la circolazione delle informazioni (Marazzi, 1994).

5. Al posto di una conclusione

A dispetto del recente successo ottenuto da metafore vitaliste delle lingue (che nascono e muoiono) sia presso organizzazioni non-profit internazionali così come aziende hi-tech californiane, per comprendere la distinzione tra lingue a "bassa" e "alta" risorsa negli odierni Modelli linguistici di grandi dimensioni, anziché postulare scenari catastrofici sull'estinzione, potremmo iniziare con la valorizzazione delle connessioni semantiche e delle catene linguistiche delineate da Virno e Marazzi. Il linguaggio in questione è quello che genera organizzazione all'interno della divisione sociale del lavoro. In questo scenario, la definizione di lingue a "bassa" (ligure) o ad "alta" (inglese) risorsa, funziona come un indice per discernere gerarchie di capacità produttiva. Allora capisco come la missione di Meta sia meno etica che speculativa.

3 *Grammatica della moltitudine*, pubblicato nel 2001, riassume i contributi di Virno in riviste e giornali del decennio precedente.

Riferimenti bibliografici

Baldwin, D. & Noodin, Margaret & Perley, B.C. 2018. Surviving the Sixth Extinction: American Indian Strategies for Life in the New World. in After Extinction, R. Grusin (ed.). Minneapolis: University of Minnesota Press.

Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell. S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? doi.org/10.1145/3442188.3445922.

Bird, S. & Chiang, D. 2012. Machine translation for language preservation. in Proceedings of the 24th International Conference on Computational Linguistics. pp. 125-134, International Conference on Computational Linguistics, Mumbai, India, 8/12/12.

Cieri, C., Maxwell, M., Strassel, S., Tracey, J. 2016. Selection criteria for low resource language programs. In Proceedings of the Tenth International Conference on Language Resources and Evaluation.

Kornai, A. 2013. Digital Language Death. PLoS ONE 8(10): e77056. doi:10.1371/journal.pone.0077056

Maffi. L. (ed.) 2001. On Biocultural Diversity: Linking Language, Knowledge, and the Environment. Washington, D.C.: Smithsonian Institution Press.

Magueresse, A., Carles, V., Heetderks, E. 2020. Low-resource languages: A review of past work and future challenges. arXiv preprint arXiv:2006.07264.

Marazzi, C. 1994. Il posto dei calzini: La svolta linguistica dell'economia e i suoi effetti nella politica. Bellinzona: Edizioni Casagrande.

Nettle, D. & Romaine, S. 2000. *Vanishing Voices: The Extinction of the World's Languages*. Oxford: Oxford University press.

NLLB Team et al. 2022. No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.

Ostler, N. 2010. The Last Lingua Franca: English Until the Return of Babel. New York: Walker & Company.

Sepkoski, D. 2020. Catastrophic Thinking: Extinction and the Value of Diversity from Darwin to the Anthropocene. Chicago: University of Chicago Press.

Strubell, E., Ganesh, A., McCallum. A. 2019. Energy and Policy Considerations for Deep Learning in NLP. in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, July 28 – August 2.

Virno, P. 2001. Grammatica della moltitudine: Per una analisi delle forme di vita contemporanee. Catanzaro: Rubettino.