MAPPE DEL LABIRINTO

UNA SERIE DI PHANTOM MAZE — AI & LANGUAGE LAB

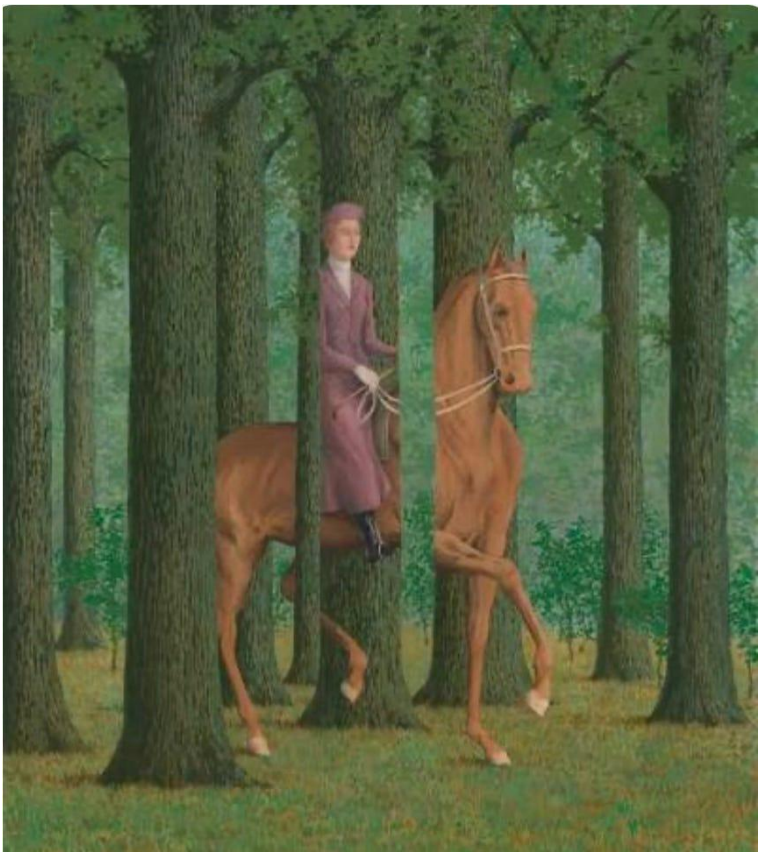
SAGGIO 19

La straordinaria borsa di Tristan Tzara

Claudia Marsico · Hernán Inverso

DOI 10.5281/zenodo.23575221

La codifica posizionale nei modelli di IA, ovvero fin dove conta l'ordine delle parole



R. Magritte, Le Blanc-Seing (1965)

Nel 1920, in pieno scandalo dadaista, Tristan Tzara pubblicò prima su *Littérature* e pochi mesi dopo nella raccolta *Manifesto sull'amore debole e l'amore amaro* una ricetta per fabbricare poesie:

Prendete un giornale.
Prendete un paio di forbici.
Scegliete nel giornale un articolo che abbia la lunghezza che volete dare alla vostra poesia.

Ritagliate l'articolo.
Ritagliate poi con cura ciascuna delle parole che formano l'articolo e mettetele in una borsa.
Agitate dolcemente.
Tirate fuori ora i ritagli uno dopo l'altro.
Copiate coscienziosamente
nell'ordine in cui sono usciti dalla borsa.
La poesia vi somiglierà.
Ed eccovi uno scrittore infinitamente originale e di una sensibilità incantevole, sebbene incompresa dal volgo.

Questa beffa dell'antiarte contro la solennità letteraria sfiora un punto curioso. Se togliamo l'ordine alle parole, il senso non sopravvive. La borsa di Tzara contiene esattamente le stesse parole dell'articolo di giornale e non dice niente di simile. Guardiamo ora il transformer, l'architettura di tutti gli attuali modelli di linguaggio. A modo suo è un lettore di borse di Tzara, dato che il suo disegno non gli lascia vedere l'ordine. Nell'elaborare i token, quei frammenti con cui opera, tutti i termini della frase vengono elaborati in una volta, ciascuno consultando gli altri e quindi senza ordine. Se diamo alla macchina "il cane ha morso il postino" e "il postino ha morso il cane", alla rinfusa, i suoi conti interni producono lo stesso risultato. I matematici dicono con eleganza che l'attenzione è invariante per permutazioni. Tzara avrebbe potuto dire che il modello legge ciò che esce dalla borsa, sicché ha bisogno che qualcuno gli restituisca l'ordine, perché non pretende di essere un dadaista. Come si insegna l'ordine a chi non ce l'ha?

R. Magritte, Le Blanc-Seing (1965)

Mettere "ordine" era in origine sistemare i fili del tessuto, ordire la trama, il che lo rende della stessa famiglia di "testo" e del suo rimando al tessere. Ma "ordine" saltò dal telaio al mondo intero, organizzandolo come fosse una tela fatta di ordini religiosi e cavallereschi, ordinazioni di sacerdoti, numeri ordinali, ordini del giorno e perfino calcolatori, che francesi e spagnoli chiamano *ordinateurs* e *ordenadores*. L'ordine delle parole è l'intelaiatura del testo, sicché la borsa di Tzara disfa la tela senza tagliare i fili.

A rigore, ci sono lingue che hanno vissuto con la borsa mezza aperta. Il latino e il greco classici avevano un ordine delle parole molto libero. L'*Eneide* comincia con *arma virumque cano*, "le armi e l'uomo canto". Nessun romano si perdeva. La poesia lo portò all'estremo con l'iperbato, la figura che disloca l'ordine atteso. Il record di eccesso glorioso spetta a Góngora, che nelle *Soledades* poteva scrivere "pasos de un peregrino son errante", "passi di un pellegrino sono errante", invece di "son pasos de un peregrino errante", "sono passi di un pellegrino errante", spargendo il sintagma lungo il verso, fiducioso che il lettore sarebbe riuscito a rimontarlo. "Pazienza devi avere", "molto da imparare ancora hai", dice il maestro Yoda nella saga di *Guerre stellari* con iperbato allo stato puro.

L'ordine è tanto più alterabile quanto più il senso dipende dalla morfologia e non dalla posizione. Il latino declina, cioè dà a ogni parola una terminazione che dice quale ruolo gioca nella frase. *Canis virum mordet* e *virum canis mordet* dicono esattamente la stessa cosa (il cane morde l'uomo) perché

“canis” porta addosso il marchio di soggetto e “virum” quello di oggetto, vada ciascuno dove vuole. Le lingue moderne, dall’inglese al cinese, hanno preferito indebolire o abolire le declinazioni e affidarsi all’ordine, e per questo la borsa di Tzara è letale per un giornale di oggi e sarebbe stata quasi innocua per Virgilio, dato che ogni ritaglio latino esce dalla borsa con la sua funzione tatuata.

Se guardiamo dalla filologia, ciò che fecero gli autori del transformer nel 2017 è sorprendentemente classico: convertirono una lingua di ordine in una lingua di declinazione. Poiché la loro macchina non sapeva leggere in fila, stamparono su ogni parola, nel corpo stesso della parola, il marchio della sua posizione. Prima di entrare in gioco, ogni token arriva convertito in un vettore, quella lista di numeri che è il suo punto nello spazio del senso. La codifica posizionale somma a quel vettore un’altra lista di numeri che codifica “sono la parola numero sette” e così il token acquista il marchio di posizione. Il disordine della borsa di Tzara non importa, perché l’ordine sta dentro ogni parola e quel filo non si aggroviglierà più.

Ogni posizione riceve la sua firma fatta di onde, in una collezione di seni e coseni di frequenze diverse, alcune che oscillano rapide e altre lentissime. Immaginiamo un orologio con moltissime lancette dove nessuna da sola dice l’ora, ma tutte insieme identificano l’istante senza ambiguità. Due istanti vicini hanno fotografie simili, due lontani fotografie molto diverse, con il che anche la distanza tra le parole resta suggerita. La firma posizionale si somma, numero per numero, al vettore del significato, sicché posizione e senso finiscono mescolati nelle stesse coordinate, come se l’accusativo non si aggiungesse alla parola ma si dissolvesse in essa. Che una somma così semplice basti a conservare insieme identità e posizione fu una di quelle scoperte in cui l’ingegneria mostrò prima che il meccanismo funzionava e la teoria dovette poi spiegare perché funzionasse tanto bene.

La copia dei libri aveva già affrontato un problema simile, che abbiamo esaminato a proposito dell’architettura transformer e dell’attenzione simultanea e che risulta significativo anche qui per il rimando di una parte all’altra. Un codice medievale si copiava per fascicoli sciolti e poi quei fogli indipendenti arrivavano al legatore. Poiché non aveva letto il testo e doveva cucirli nell’ordine giusto, si stampavano marchi d’ordine su ogni pezzo per mezzo delle *segnature*, lettere e numeri in calce ai fascicoli, e del *richiamo*, che indicava la prima parola del fascicolo successivo. La codifica posizionale è il richiamo del token, cioè il marchio che permette a un’elaborazione cieca e simultanea di ricostruire una sequenza che non ha mai letto in quel modo.

La firma di onde del 2017 fu solo il primo modo di risolvere il problema. Poco dopo, modelli come GPT-2 e BERT cominciarono a imparare anche la posizione. Invece di ricevere una firma matematica già disegnata per “posizione 1”, “posizione 2” o “posizione 3”, il modello impara un vettore diverso per ogni posto della sequenza. Questo gli dà più libertà, con lo svantaggio che conosce bene le posizioni per cui è stato addestrato e non sa che fare quando il testo va oltre. Che cosa fa un legatore che ha imparato a numerare fino a un certo numero di pagine quando all’improvviso riceve un volume molto più lungo?

Nel 2018 sorse l’idea di concentrarsi sulla relazione tra le parole più che sul loro numero assoluto. Non conta tanto che “cane” stia in posizione sette quanto che stia tre posti dopo “il” o due prima di

“abbaia”. La rete poteva imparare quanto separava certe parole da altre. A rigore, la grammatica è fatta più di vicinanze e distanze che di numeri fissi. Da quell’idea nacque RoPE, “Rotary Position Embedding”, proposto nel 2021, dove la posizione smette di sommarsi come una seconda etichetta numerica e si incorpora facendo ruotare la rappresentazione di ogni parola secondo il posto che occupa. La parola continua a essere rappresentata da un insieme di numeri, ma ora ruotati in modo leggermente diverso a seconda che compaia per prima, quinta o centesima. La differenza tra quelle rotazioni permette di recuperare informazione su quanto separa determinate parole, come se ogni parola portasse una lancetta d’orologio che fa sapere dove sta rispetto alle altre.

Una soluzione ancora più austera fu ALiBi, presentata nel 2022, che invece di stampare la posizione sulle parole agisce direttamente sull’attenzione tra di esse: quanto più una è lontana dall’altra, tanto più la relazione viene penalizzata, sicché, a parità di altre condizioni, le parole vicine conservano più influenza di quelle remote senza bisogno di imparare un marchio specifico per ogni posizione. La semplicità del procedimento aveva inoltre un vantaggio considerevole, riassunto nel titolo del paper, “Train Short, Test Long”: un modello addestrato con sequenze relativamente brevi poteva poi estendersi a testi più lunghi senza perdersi in un sistema di posizioni che non aveva imparato. Questo dettaglio ha finito per diventare un campo di battaglia. Il modo in cui il modello rappresenta la posizione condiziona fin dove può seguire il filo di un testo prima di restare senza riferimenti utili. Da questo dipende, in buona misura, la “finestra di contesto” che le aziende annunciano come fosse la cilindrata di un motore e che cresce a ogni miglioria.

La ricetta di Tzara finiva con una promessa ironica: dopo aver ritagliato le parole, averle messe in una borsa e averle tirate fuori a caso, “la poesia vi somiglierà”. Il modello di linguaggio crea una sensazione simile. Quante persone sentono che il modello le conosce? È addestrato a rispondere in modi che ci risultino attraenti e quanto meglio lo fa, tanto più è facile confondere quell’adattamento con la somiglianza. La risposta può adottare il nostro stile e anticipare le nostre aspettative, e sotto quella familiarità resta la borsa di Tzara con i suoi pezzi discreti, posizioni, probabilità e relazioni numeriche. La somiglianza del risultato non dovrebbe farci dimenticare le differenze.

COME CITARE

Marsico, Claudia e Hernán Inverso. 2026. “La straordinaria borsa di Tristan Tzara”. *Mappe del labirinto*, saggio 19. Phantom Maze — AI & Language Lab. https://doi.org/10.5281/zenodo.23575221

Mappe del labirinto è una serie di Phantom Maze, AI & Language Lab. La collezione vive su **phantommaze.cloud**. L'iscrizione è gratuita.

Alcune delle idee che inquadrano questo testo sono sviluppate in *Il fantasma senza ego. Una mappa nel labirinto dell'intelligenza artificiale*, di Hernán Inverso e Claudia Marsico (Phantom Maze Press, 2026; ISBN 978-987-3729-19-5).

Fonti. Ashish Vaswani et al., “Attention Is All You Need” (NeurIPS, 2017), sezione 3.5; Jacob Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers” (2018); Alec Radford et al., “Language Models are Unsupervised Multitask Learners” (2019); Peter Shaw, Jakob Uszkoreit e Ashish Vaswani, “Self-Attention with Relative Position Representations” (NAACL, 2018); Jianlin Su et al., “RoFormer: Enhanced Transformer with Rotary Position Embedding” (2021); Ofir Press, Noah A. Smith e Mike Lewis, “Train Short, Test Long: Attention with Linear Biases Enables Input Length Generalization” (ICLR, 2022).