CRS

Centro per la Riforma dello Stato

LE NUOVE FABBRICHE DEL LINGUAGGIO

COME FUNZIONANO, CHI LE POSSIEDE, COME OCCUPARLE

Guido Vetere

15/11/21

FACEBOOK IN MYANMAR, 2018

«La velocità con cui vengono segnalati contenuti inappropriati [bad] *in birmano, che si tratti di incitamento all'odio o disinformazione, è bassa.* […] ***Quindi stiamo investendo molto nell'intelligenza artificiale in grado di segnalare in modo proattivo i post che infrangono le nostre regole»***

https://about.fb.com/news/2018/08/update-on-myanmar/

LINGUAGGIO E INTELLIGENZA ARTIFICIALE

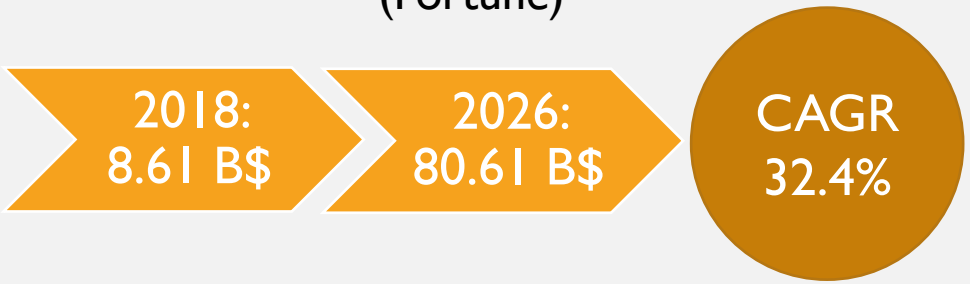
Il mezzo è il messaggio:

- XX sec: è = definisce le condizioni
- XXI sec: è = definisce i contenuti

L'INDUSTRIA DEL LINGUAGGIO

Ricerca di informazione	• Motori di ricerca • Gestione dei contenuti • Business intelligence
Traduzione automatica	• Commercio elettronico • Editoria • Relazioni istituzionali
Analisi e classificazione	• Sicurezza • Reti sociali • Pubblicità
Interazione	• Supporto alla clientela • Relazioni col pubblico • Assistenza personale
Generazione di testi	• Gioco e intrattenimento • Editoria • Lavoro creativo

Global Natural Language Processing Market (Fortune)



Global IT Services Market (Research and Markets)

USI SOCIALI DELLE TECNOLOGIE LINGUISTICHE (ESEMPIO)

USI PERVERSI DELLE TECNOLOGIE LINGUISTICHE

Microsoft sacks journalists to replace them with robots

Users of the homepages of the MSN website and Edge browser will now see news stories generated by AI

From July, the MSN homepage will no longer feature news stories produced by journalists at PA Media, formerly the Press Association. Photograph: Alamy

Dozens of journalists have been sacked after Microsoft decided to replace them with artificial intelligence software.

The Guardian

Jim Waterson, 30 Maggio 2020

www.theguardian.com/technology/2020/may/30/microsoft-sacks-journalists-to-replace-them-with-robots

POSSIBILI USI DELLE TECNOLOGIE LINGUISTICHE ANCORA PIÙ PERVERSI

CNET Highlights
212.000 iscritti

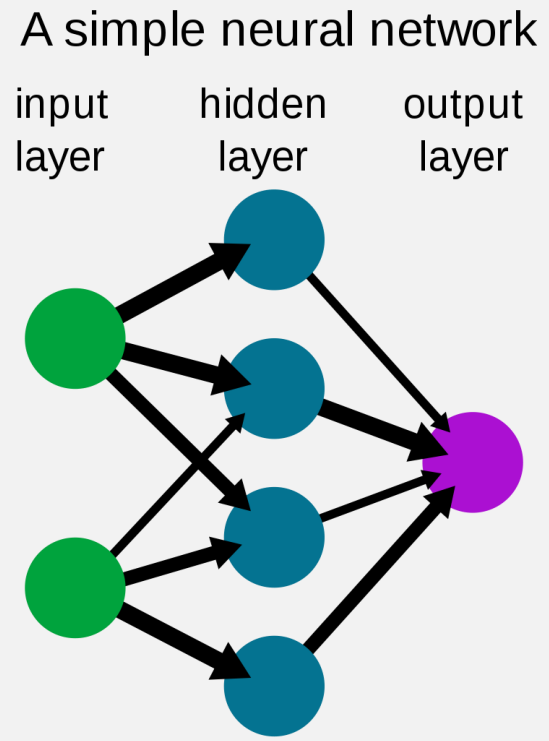
28/10/21: A Facebook Connect, il CEO Mark Zuckerberg svela la visione della sua azienda su come gli utenti interagiranno e socializzeranno all'interno del Metaverso.

https://www.youtube.com/watch?v=b9vWShsmE20

TECNOLOGIE LINGUISTICHE E INTELLIGENZA ARTIFICIALE

Le tecnologie della lingua (NLP) si basano oggi principalmente sulle capacità di apprendimento delle «reti neurali»

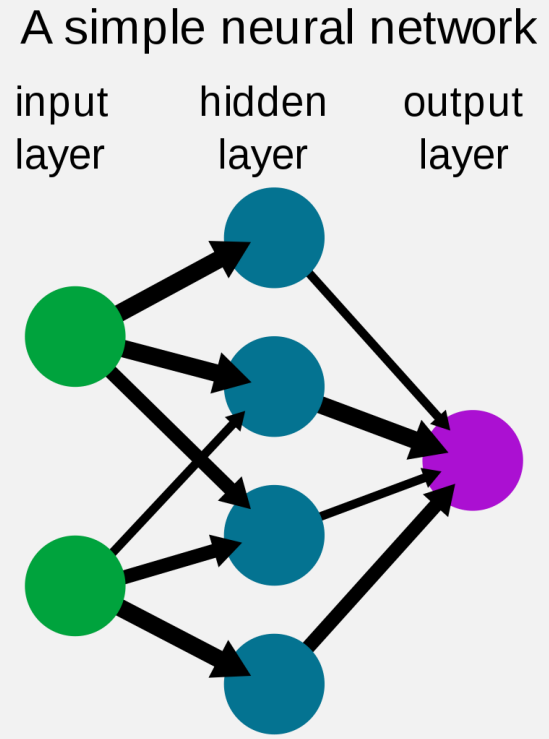
- Le reti neurali sono **classificatori** addestrati su campioni di dati (*training set*)
  - «Imparano» a produrre un *output* a fronte di un *input* mediante il rinforzo (o l'indebolimento) delle connessioni tra elementi atomici detti «neuroni»
  - Richiedono una specifica codifica dei dati, una complessa architettura interna, e una funzione di calcolo dell'errore nella predizione del risultato (*loss function*)



TECNOLOGIE LINGUISTICHE E INTELLIGENZA ARTIFICIALE

Vi sono diversi modi per addestrare le «reti neurali», i quali rientrano in due classi:

- Apprendimento supervisionato
  - Il *training set* è annotato da esseri umani, i quali forniscono esempi (positivi e\o negativi) di quale *output* si attende a fronte di quale *input*
- Apprendimento non supervisionato
  - Il *training set* è costituito da dati «grezzi», l'associazione tra *input* e *output* è intrinseca

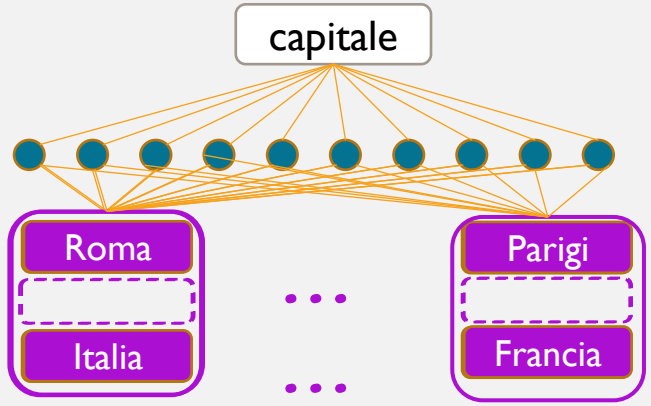


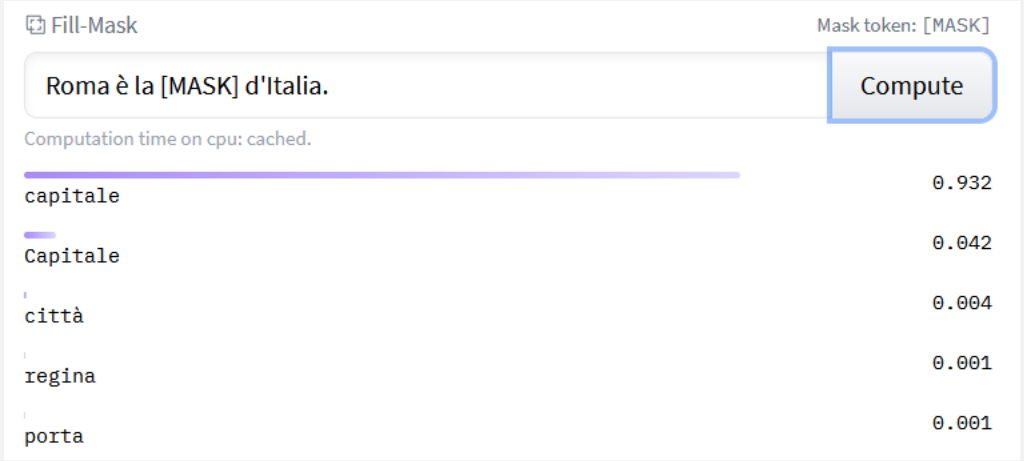
Il salto di qualità delle tecnologie linguistiche è dovuto principalmente all'apprendimento non supervisionato

MODELLAZIONE LINGUISTICA NON SUPERVISIONATA

Le unità del linguaggio sono rappresentate numericamente

- Per ogni token (parola o frammento), una rete apprende a predire il contesto in cui appare (es. *skip-gram model*)
- Dalle connessioni (parametri) della rete addestrata vengono estratti vettori numerici chiamati *embedding*
- L'insieme degli *embedding* stimati sui testi di una lingua è chiamato «modello linguistico» (*language model*)





https://huggingface.co/

capitale	0.5	0.1	0.9	0.7	0.0	0.0	0.6	0.3	1.0	.
città	0.4	0.2	0.9	0.5	0.1	0.0	0.5	0.5	0.9	.

USO DEI MODELLI LINGUISTICI

downstream *task*

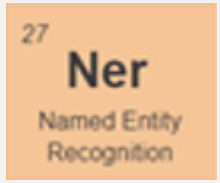
Modelli generali
- Testi
- Apprendimento non supervisionato
- Validazione sperimentale

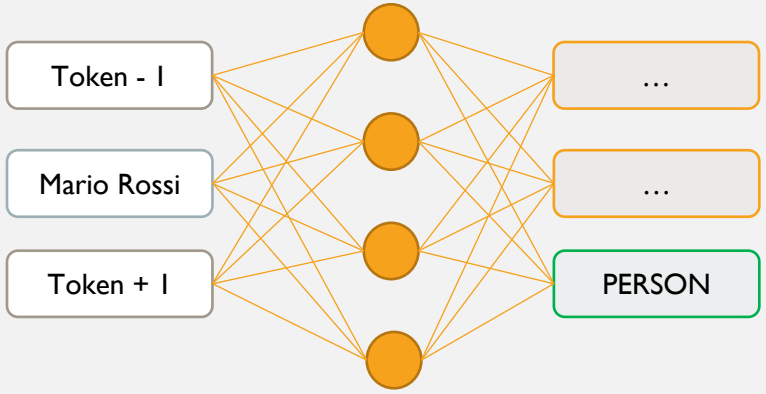
Modelli specifici
- Annotazioni
- Apprendimento supervisionato
- Validazione sperimentale

Task
- Integrazione
- Validazione applicativa

R. van Zoest: 6 NLP Tasks for Training Data Generation (medium.com)

ANALISI DI RIFERIMENTI E CONCETTI (*NAMED ENTITY RECOGNITION*)





- Le reti sono addestrate per classificare ciascun «token» sulla base del contesto in cui appare
- L'addestramento è supervisionato mediante corpora annotati e si avvale normalmente di modelli linguistici
- I corpora annotati si possono ottenere tramite *crowdsourcing*, o sfruttando risorse esistenti (es. Wikipedia)

Il numero delle etichette è in genere molto limitato (<20)

USO DEI MODELLI LINGUISTICI

zero-shot

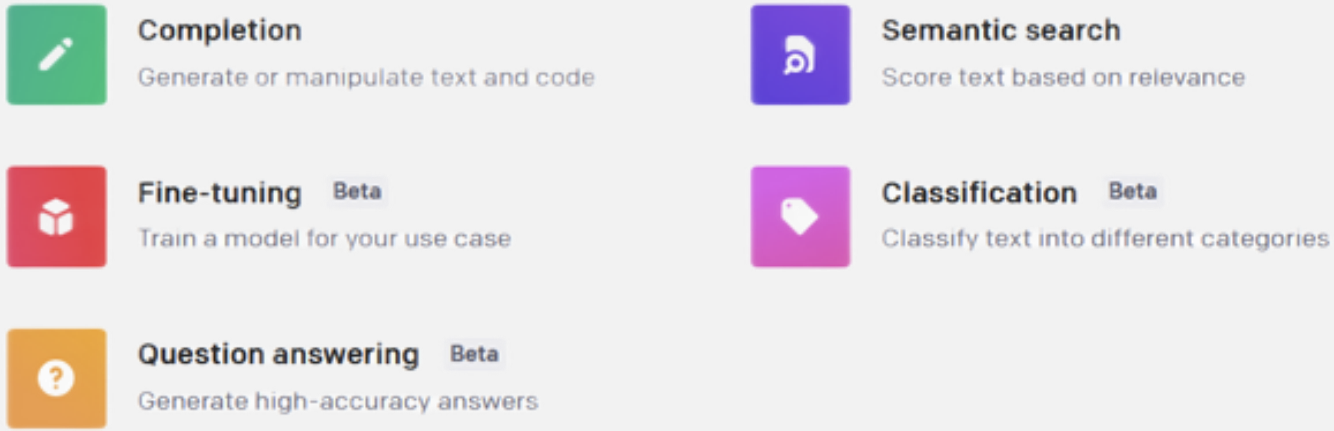
Modelli generali

- Infosfera (multilingue)
- Apprendimento non supervisionato

Generatori

- *Pre-trained transformer*

OpenAI *GPT-3*



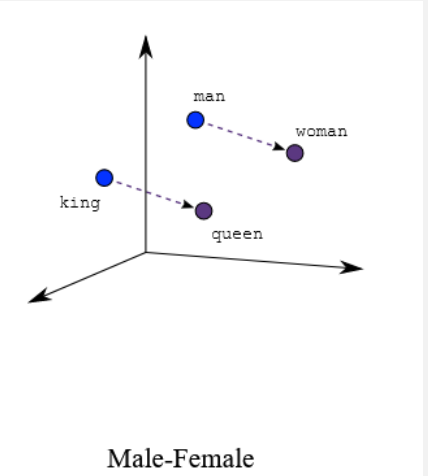
RETE NEURALE: 175 MLD PARAMETRI
COSTI SVILUPPO: 11 < M$ > 27 (STIME)
ENERGIA: 190K kWh (85T CO_2)

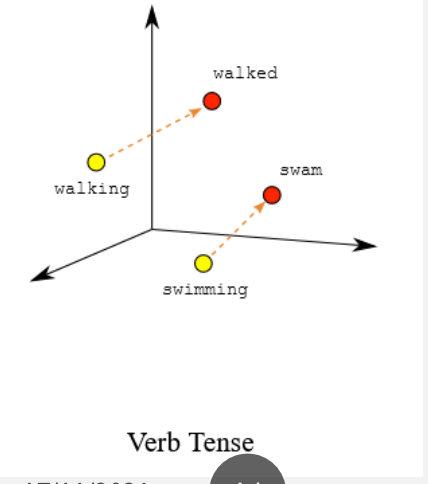
B. Dickson, The GPT-3 economy
https://bdtechtalks.com/2020/09/21/gpt-3-economy-business-model/

LA SEMANTICA DEI MODELLI NEURALI

I modelli linguistici neurali considerano il significato in modo **distribuzionale**

- Gli *embedding* di parole che compaiono negli stessi contesti risultano simili
- I loro vettori vengono mappati in uno «spazio» n-dimensionale
- Le relazioni semantiche sono viste come misure di distanza in questo spazio





"Sia "cane" che "gatto" appaiono spesso vicini alla parola "veterinario", e questo fatto riflette la loro somiglianza semantica" (Google, 2020)

"You shall know a word by the company it keeps" (Firth, 1957)

LA SEMANTICA DEI MODELLI NEURALI

I modelli distribuzionali hanno accesso solo al «significante» del segno linguistico

- La «semantica» delle reti neurali fa emergere solo le relazioni concettuali che hanno effetto sulla distribuzione delle parole nei testi forniti
- Ciò che è fuori dal testo (*hors-texte*) per l'automa non esiste
- Più in generale, non esiste, per l'automa, alcun **processo di significazione**
- Denotazione e connotazione dunque collassano, da cui (tra l'altro) il noto problema dei pregiudizi (*bias*)

ARCHITETTURE E MODELLI DI BUSINESS

cloud

- Software e modelli linguistici gestiti nei server dei fornitori
- Licenze per l'uso remoto (API)
- Funzionalità *out-of-the-box*
- *Vendor-lock, protezione dati*

on-premise

- Software e modelli linguistici gestiti dagli utilizzatori

Sviluppo in-house

- Open source
- Modelli aperti
- *Competenze*

Terze parti

- Piattaforme Enterprise
- Modelli custom
- *Vendor-lock*

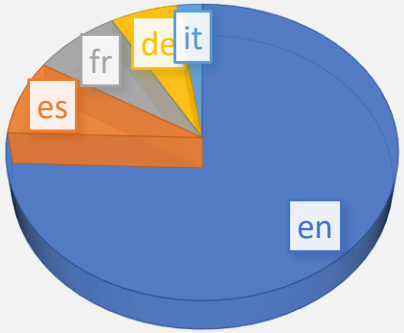
Centralizzazione ● Monopolio ● *Knowledge divide*

Decentralizzazione ● Mercato ● *Knowledge commons*

LINGUE DIVERSE DALL'INGLESE

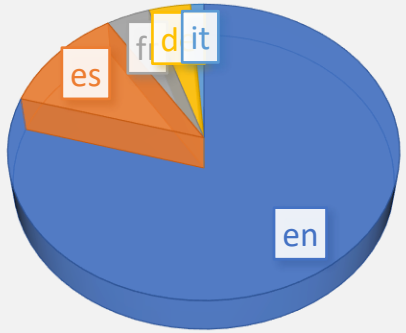
Hugging Face

Lingua	Modelli	Dataset
en	2486	309
es	269	72
fr	261	68
de	199	68
it	76	48



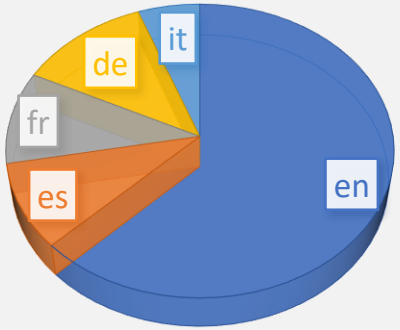
LDC Linguistic Data Consortium

Lingua	Risorse
en	524
es	77
fr	26
de	25
it	8



EUROPEAN LANGUAGE GRID

Lingua	Corpora	Onto\Lex
en	1535	464
es	229	298
fr	250	144
de	295	121
it	139	68



- Le tecnologie linguistiche dipendono oggi dalla disponibilità di dati grezzi o pre-elaborati (modelli)
- Tra i temi del *knowledge divide* c'è lo squilibrio nella produzione di risorse linguistico-computazionali

META

Studio completo (2012)
http://www.meta-net.eu

LINGUE DIVERSE DALL'INGLESE

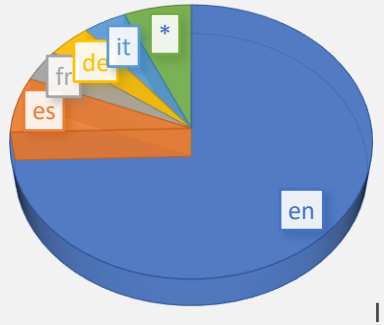
Il «paradigma open» non risolve di per sé il problema del predominio dell'inglese nelle tecnologie linguistiche

Spark NLP models & pipelines	en	es	fr	de	it	*
Named Entities	38	8	5	5	5	
Text Classification	15	0	0	0	0	
Sentiment Analysis	5	0	0	0	0	
Translation	3	0	(1)	(1)	0	
Question Answering	5	0	0	0	0	
Summarization	4	0	0	0	0	
Sentence Detection	1	0	0	0	0	9
Embeddings	101	0	0	0	0	10
Part-of-speech	7	5	2	3	3	
Lemmatization	19	6	3	3	3	
Relation Extraction	12	0	0	0	0	
Spell check	4	0	0	0	1	
Totale	214	19	11	12	12	19

John Snow LABS

Spark NLP è la piattaforma aperta più usata per scopi industriali (Gradient Flow: NLP Industry Survey, 2020)

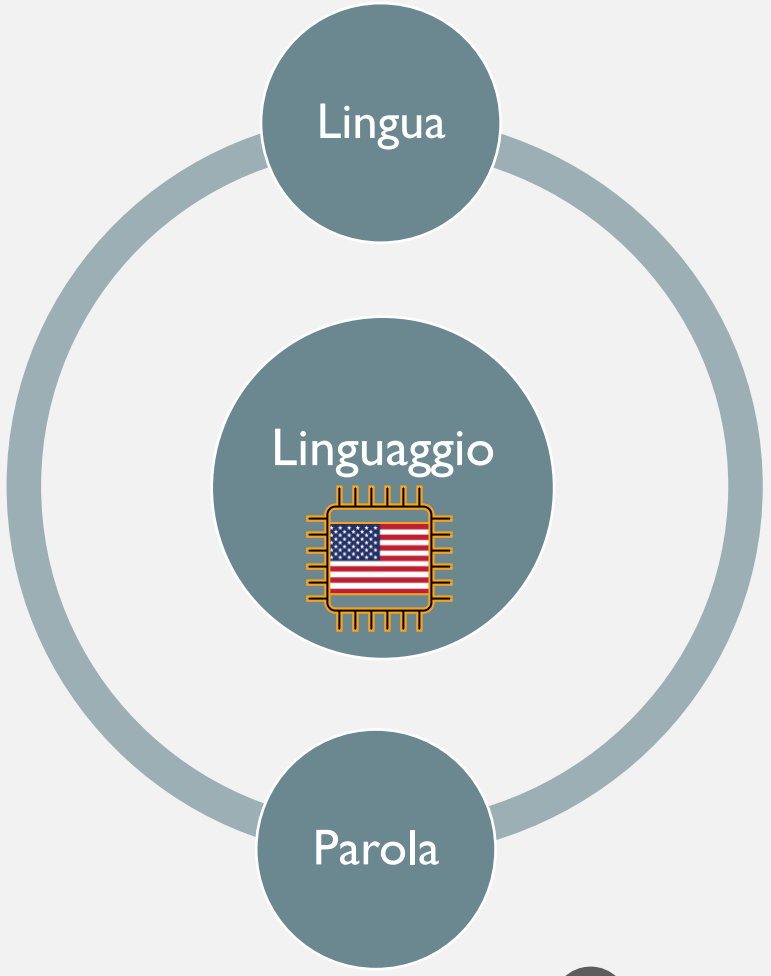
- Basato su Apache Spark
- Programmabile in Python e Scala
- 1426 modelli e *pipeline* «out of the box»



L'INGLESE DA *LINGUA* A *LINGUAGGIO*

- L'AI anglofona rischia oggi di ridefinire la **facoltà del linguaggio** su scala globale
- Non più una lingua tra le altre, ma il presupposto stesso della vita linguistica

Colonizzazione	Omologazione
• Trasferimento di risorse verso i monopoli digitali • Influenza nelle economie locali	• Pressione dell'inglese sulle lingue «di minoranza» • Perdita di diversità linguistica \ culturale

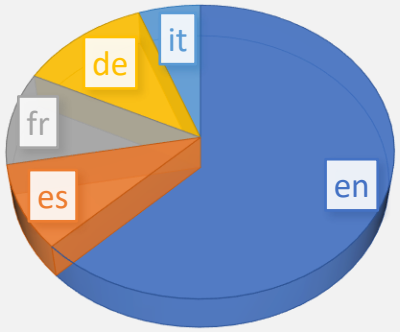


COSA STA FACENDO L'EUROPA

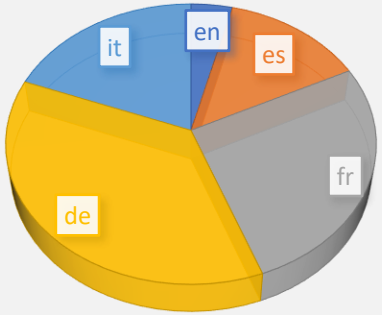
EUROPEAN LANGUAGE EQUALITY

Developing an agenda and a roadmap for achieving full digital language equality in Europe by 2030

MULTILINGUAL EUROPE TECHNOLOGY ALLIANCE



10 anni



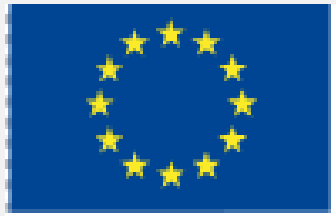
Lingua	PIL 2020
en	306
es	1058
fr	2061
de	2832
it	1479

GDP 2020 per area linguistica delle maggiori economie EU + l'Irlanda

EUROPEAN LANGUAGE GRID

Towards the Primary Platform for Language Technologies in Europe

COSA STA FACENDO L'EUROPA



21\4\21: Proposta per una regolamentazione dell'approccio europeo all'Intelligenza Artificiale

Di fronte al rapido sviluppo tecnologico dell'IA e a un contesto politico globale in cui sempre più paesi investono fortemente nell'IA, l'UE deve agire unitariamente per sfruttare le numerose opportunità e affrontare le sfide dell'IA in modo adeguato agli sviluppi futuri.

https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence

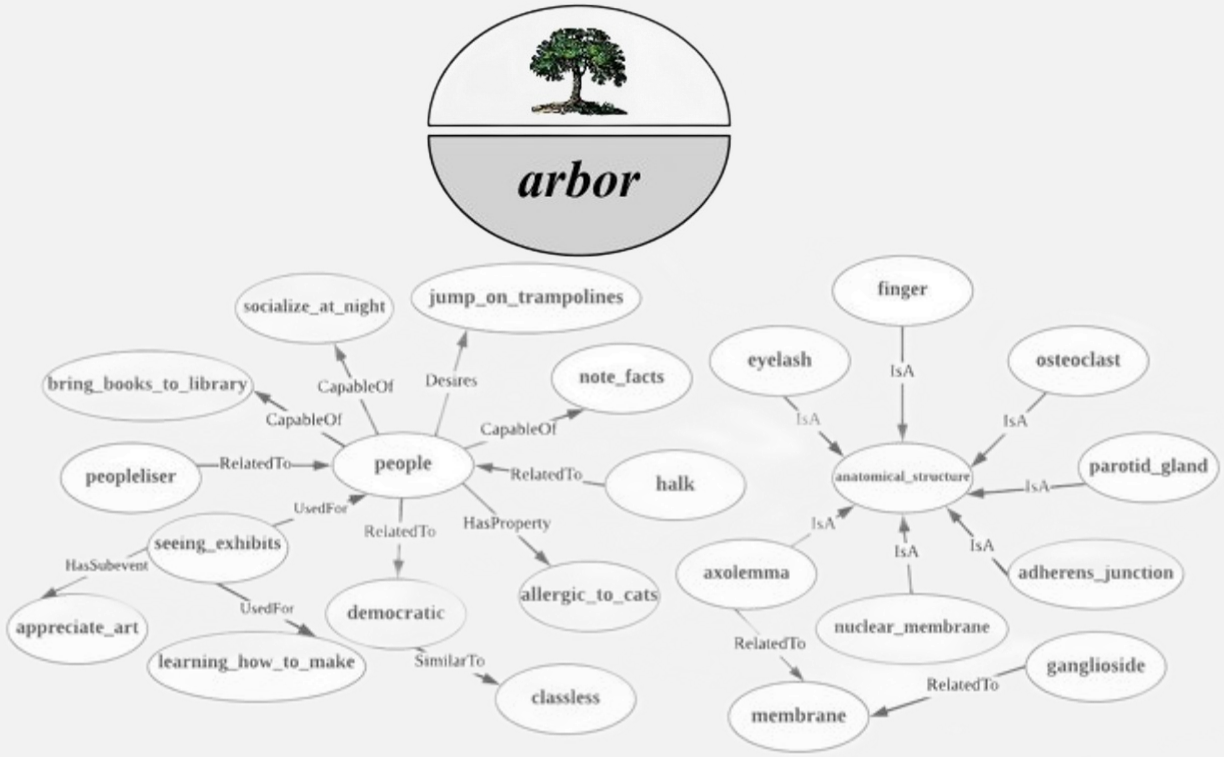
- Trasparenza e verificabilità dei dati di addestramento
- Considerazione della specificità delle diverse culture
- Eliminazione dei pregiudizi (bias)
- Trasparenza dei processi di generazione dei risultati (auditability)
- Contrasto agli usi manipolativi

PROSPETTIVE DI RICERCA: BENTORNATO SEGNO!

Gli *embedding* ottenibili da risorse lessico \ ontologiche (aka Knowledge Graph, es. WordNet, ConceptNet, BabelNet) sono competitivi per molti *task*

- Controllo sociale, plasmabilità
- Sostenibilità eco-tecnologica
- Supporto a lingue in scarsità di risorse
- Supporto a lingue morfologicamente complesse

Anche sui Knowledge Graph , tuttavia, c'è da colmare un grande divario tra l'inglese e le altre lingue

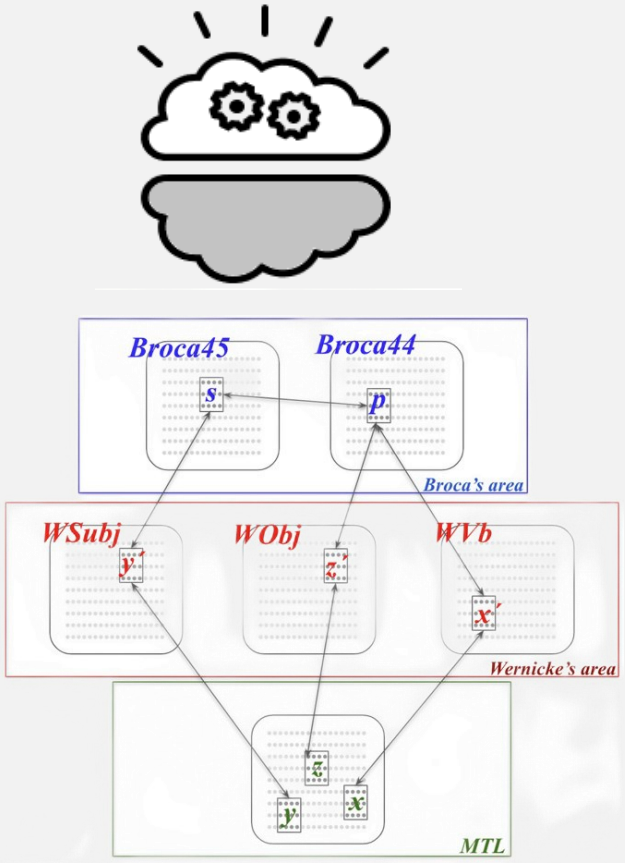


ConceptNet Knowledge Graph
A Data-Driven Study of Commonsense Knowledge using the ConceptNet Knowledge Base, Ke Shen, Mayank Kejriwal, 2020

PROSPETTIVE DI RICERCA: BENTORNATO PENSIERO!

- Modellazione **plausibile** delle strutture neurali che sono alla base delle funzioni cognitive complesse
- Assembly Calculus: proiezione, associazione, fusione
- *Pattern completion*
- Principale campo applicativo: il linguaggio (primi esperimenti con la sintassi già in corso)

Linguaggio come fondamento della cognitività umana nelle pratiche sociali (Lev S. Vygotskij, *Pensiero e linguaggio,* Laterza 1990)



Brain computation by assemblies of neurons
Christos H. Papadimitriou et al, 2020

CONCLUSIONE

- L'AI moderna (reti neurali) ha rivoluzionato le tecnologie linguistiche e ne ha aumentato significativamente la portata
- Tuttavia, molte attuali tecnologie sradicano il linguaggio dai processi di significazione che costituiscono il loro fondamento sociale
- Inoltre, per come sono prodotte e usate, tali tecnologie aumentano le concentrazioni e il divario in favore delle economie anglofone
- Questo divario esercita una pressione sulla vita linguistica dei Paesi europei
- Per favorire uno sviluppo equilibrato delle tecnologie linguistiche, l'Europa dovrà diffondere competenze e investire risorse (regolamenti e «marketplace» non bastano)
- La ricerca di modelli alternativi a quelli basati sulla «forza bruta» delle reti neurali è promettente e va sviluppata