

Le risorse linguistiche al tempo delle reti neurali



Seminario di Cultura Digitale, 8/03/2023

Alessio Miaschi

ItaliaNLP Lab, Istituto di Linguistica Computazionale (CNR-ILC), Pisa

alessio.miaschi@ilc.cnr.it

<https://alemmaschi.github.io/>

<http://www.italianlp.it/alessio-miaschi/>

About me and...



Sono un PostDoc e lavoro presso l'[ItaliaNLP Lab](#), Istituto di Linguistica Computazionale “A. Zampolli” ([CNR-ILC](#), Pisa). Nel 2022 ho conseguito il dottorato di ricerca in Informatica presso l'Università di Pisa.

I miei interessi di ricerca si collocano principalmente nel contesto dell'NLP. Sono particolarmente interessato all'analisi e alla definizione di metodi per l'interpretabilità dei più recenti modelli neurali del linguaggio, nonché allo sviluppo di strumenti NLP per la realizzazione di applicazioni didattiche.

About me and... the team!



Sono un PostDoc e lavoro presso l'[ItaliaNLP Lab](http://www.italianlp.it), Istituto di Linguistica Computazionale “A. Zampolli” ([CNR-ILC](http://www.cnr-ilc.it), Pisa). Nel 2022 ho conseguito il dottorato di ricerca in Informatica presso l'Università di Pisa.

I miei interessi di ricerca si collocano principalmente nel contesto dell'NLP. Sono particolarmente interessato all'analisi e alla definizione di metodi per l'interpretabilità dei più recenti modelli neurali del linguaggio, nonché allo sviluppo di strumenti NLP per la realizzazione di applicazioni didattiche.



L'**ItaliaNLP Lab (CNR-ILC)** riunisce ricercatori, postdoc e studenti di Linguistica Computazionale, Informatica e Linguistica che lavorano allo sviluppo di risorse e algoritmi per l'elaborazione e la comprensione del linguaggio naturale.

Ricercatori:

- Felice Dell'Orletta
- Simonetta Montemagni
- Dominique Brunato
- Franco Alberto Cardillo
- Giulia Venturi

Postdocs:

- Chiara Alzetta
- Alessio Miaschi
- Andrea Amelio Ravelli

PhD Students:

- Luca Bacco
- Benedetta Iavarone
- Irene Dini
- Giovanni Puccetti

Master/Undergraduate/Visiting Students

Link al sito: <http://www.italianlp.it/>

Introduzione

AL

Generami titolo e abstract di una tesi di laurea magistrale del corso di laurea in Informatica Umanistica.



Titolo: Analisi dei modelli di apprendimento automatico per la classificazione di documenti storici  

Abstract: In questo lavoro di tesi, vengono analizzati diversi modelli di apprendimento automatico per la classificazione di documenti storici. L'obiettivo è quello di trovare un modello di classificazione efficiente ed efficace per gestire grandi volumi di dati storici digitalizzati. In particolare, ci concentriamo su documenti storici provenienti da biblioteche e archivi, in cui la classificazione accurata e veloce può migliorare notevolmente l'accesso ai dati storici da parte dei ricercatori e del pubblico. Vengono esplorate diverse tecniche di pre-elaborazione dei dati e vari algoritmi di apprendimento automatico, tra cui support vector machine (SVM), reti neurali e decision tree. Sperimentazioni sistematiche vengono condotte su un corpus di documenti storici digitalizzati per valutare le prestazioni dei vari modelli di classificazione. I risultati mostrano che il modello di apprendimento automatico basato su SVM con kernel lineare produce la migliore precisione e recall per la classificazione dei documenti storici.

AL

Generami titolo e abstract di una tesi di laurea in Informatica Umanistica.



Titolo: Analisi dei modelli di apprendimento storico

Abstract: In questo lavoro di tesi, vengono analizzati i modelli di apprendimento automatico per la classificazione di documenti storici. L'obiettivo è sviluppare un modello di classificazione efficiente ed efficace per gestire grandi volumi di dati digitalizzati. In particolare, ci concentriamo su documenti storici provenienti da biblioteche e archivi, in cui la classificazione accurata e veloce può migliorare notevolmente l'accesso ai dati storici da parte dei ricercatori e del pubblico. Vengono esplorate diverse tecniche di pre-elaborazione dei dati e vari algoritmi di apprendimento automatico, tra cui support vector machine (SVM), reti neurali e decision tree. Sperimentazioni sistematiche vengono condotte su un corpus di documenti storici digitalizzati per valutare le prestazioni dei vari modelli di classificazione. I risultati mostrano che il modello di apprendimento automatico basato su SVM con kernel lineare produce la migliore precisione e recall per la classificazione dei documenti storici.

Se ChatGpt si innamora (in modo inquietante) di chi gli fa le domande

Di Rossana Miranda | 17/02/2023 - Verde e blu



Moglie lascia il marito per l'amante su consiglio di una ChatGPT: «Mi ha dato la spinta di cui avevo bisogno»

La donna aveva una relazione extraconiugale da sei mesi mentre da cinque anni era sposata

TECNOLOGIA > NEWS
Sabato 18 Febbraio 2023



tract di una tesi di laurea in

elli di apprendime

oro di tesi, vengono analizzati

sificazione di documenti storici. L'obiettivo è

one efficiente ed efficace per gestire grandi volumi di dati

olare, ci concentriamo su documenti storici provenienti da

n cui la classificazione accurata e veloce può migliorare

notevolmente l'accesso ai dati storici da parte dei ricercatori e del pubblico. Vengono esplorate diverse tecniche di pre-elaborazione dei dati e vari algoritmi di apprendimento automatico, tra cui support vector machine (SVM), reti neurali e decision tree. Sperimentazioni sistematiche vengono condotte su un corpus di documenti storici digitalizzati per valutare le prestazioni dei vari modelli di classificazione. I risultati mostrano che il modello di apprendimento automatico basato su SVM con kernel lineare produce la migliore precisione e recall per la classificazione dei documenti storici.

Se ChatGpt si
innamora (in
modo
inquietante) di
chi gli fa le
domande

Di Rossana Miranda | 17/02/2023 - Verde e blu



Moglie lascia il marito per l'amante su consiglio di una ChatGPT: «Mi ha dato la spinta di cui avevo bisogno»

La donna aveva una relazione extraconiugale da sei mesi mentre da cinque anni era sposata

TECNOLOGIA > NEWS
Sabato 18 Febbraio 2023



notevolmente l'accesso ai dati: si possono esplorare diverse tecniche di prompt engineering automatico, tra cui support vector machines. Sperimentazioni sistematiche vengono digitalizzati per valutare le prestazioni. I risultati mostrano che il modello di apprendimento produce la migliore precisione e

tract di una tesi di laurea in
a.
elli di apprendime

oro di tesi, vengono analizzati e
sificazione di documenti storici. L'obiettivo è
one efficiente ed efficace per gestire grandi volumi di dati.
olare, ci concentriamo su documenti
n cui la cl

INTELLIGENZA ARTIFICIALE

È ChatGPT-mania su LinkedIn: ecco come l'AI può aiutare a gestire il profilo

Home > Cultura E Società Digitali



Se ChatGpt si innamora (in modo inquietante) di chi gli fa le domande

Di Rossana Miranda | 17/02/2023 - Verde e blu



Moglie lascia il marito per l'amante su consiglio di una ChatGPT: «Mi ha dato la spinta di cui avevo bisogno»

La donna aveva una relazione extraconiugale da sei mesi mentre da cinque anni era sposata

TECNOLOGIA > NEWS
Sabato 18 Febbraio 2023



CNET Your guide to a better future

Tech > Computing

Why We're Obsessed With the Mind-Blowing ChatGPT AI Chatbot

This artificial intelligence bot can answer questions, write essays, summarize documents and write software. But deep down, it doesn't know what's true.

13 min read

Stephen Shankland
Feb. 19, 2023 5:00 a.m. PT

tract di una tesi di laurea

a.

elli di apprendime

oro di tesi, vengono analizzati

ificazione di documenti storici. L'obiettivo è

efficiente ed efficace per gestire grandi volumi di dati

ci concentriamo su documenti

la cl **INTELLIGENZA ARTIFICIALE**

ati:

i pr

ect

ve

st

re

È ChatGPT-mania su LinkedIn: ecco come l'AI può aiutare a gestire il profilo

Home > Cultura E Società Digitali



Se ChatGpt si innamora (in modo inquietante) di chi gli fa le domande

Di Rossana Miranda | 17/02/2023 - Verde e blu



Moglie lascia il marito per l'amante su consiglio di una ChatGPT: «Mi ha spinto di cui avevo bisogno»

La donna aveva una relazione extraconiugale da sei mesi e cinque anni era sposata

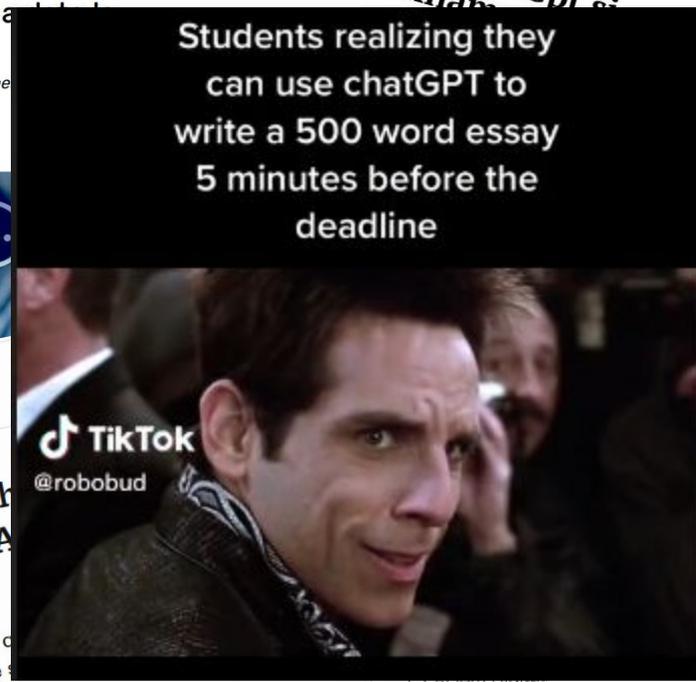
TECNOLOGIA > NEWS
Sabato 18 Febbraio 2023



tract di una tesi di laurea

Se ChatGPT è innanzi

laurea in



CNET Your guide to a better future

Tech > Computing

Why We're Obsessed With Mind-Blowing ChatGPT AI Chatbot

This artificial intelligence bot can answer questions, write essays, summarize documents and write code. But, for all its power, it doesn't know what's true.

Stephen Shankland
Feb. 19, 2023 5:00 a.m. PT

su LinkedIn: ecco come gestire il

Home > Cultura E Società Digitali



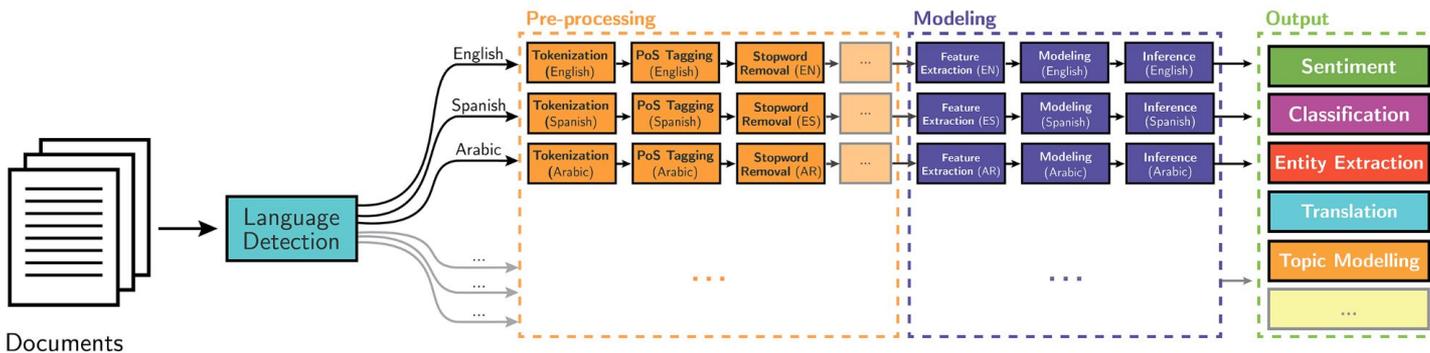
Introduzione

- Il settore dell’NLP ha visto un progresso senza precedenti nel corso degli ultimi anni
- Gran parte di questi progressi sono dovuti alla sostituzione dei “sistemi tradizionali” con più recenti modelli basati su algoritmi di Deep Learning

Introduzione

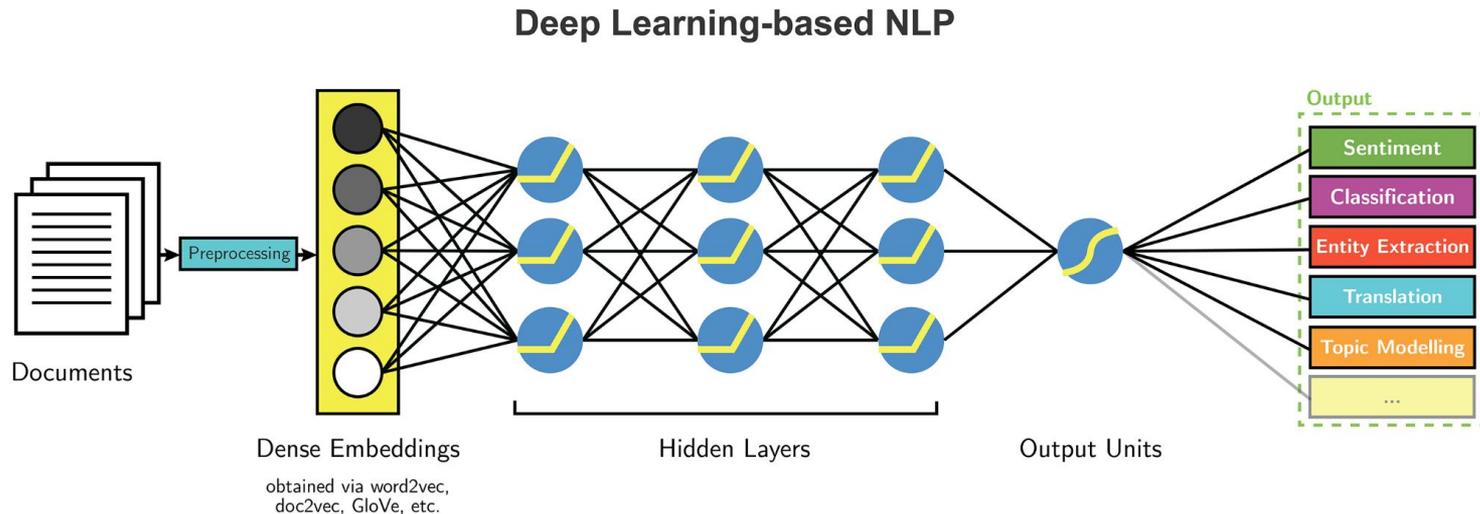
- Il settore dell’NLP ha visto un progresso senza precedenti nel corso degli ultimi anni
- Gran parte di questi progressi sono dovuti alla sostituzione dei “sistemi tradizionali” con più recenti modelli basati su algoritmi di Deep Learning

Classical NLP

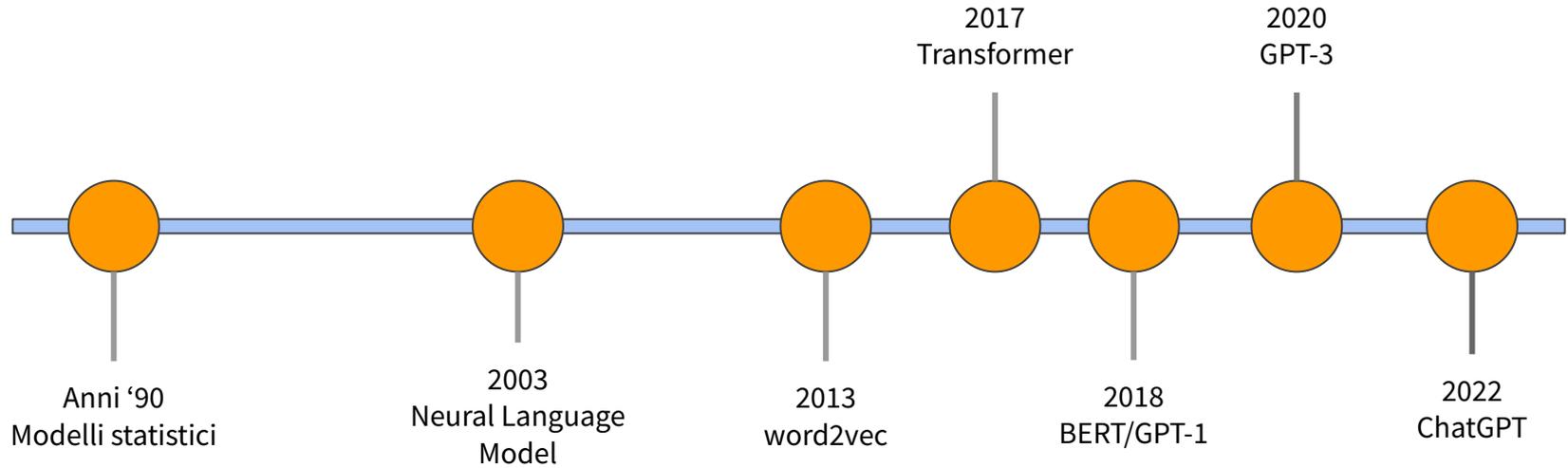


Introduzione

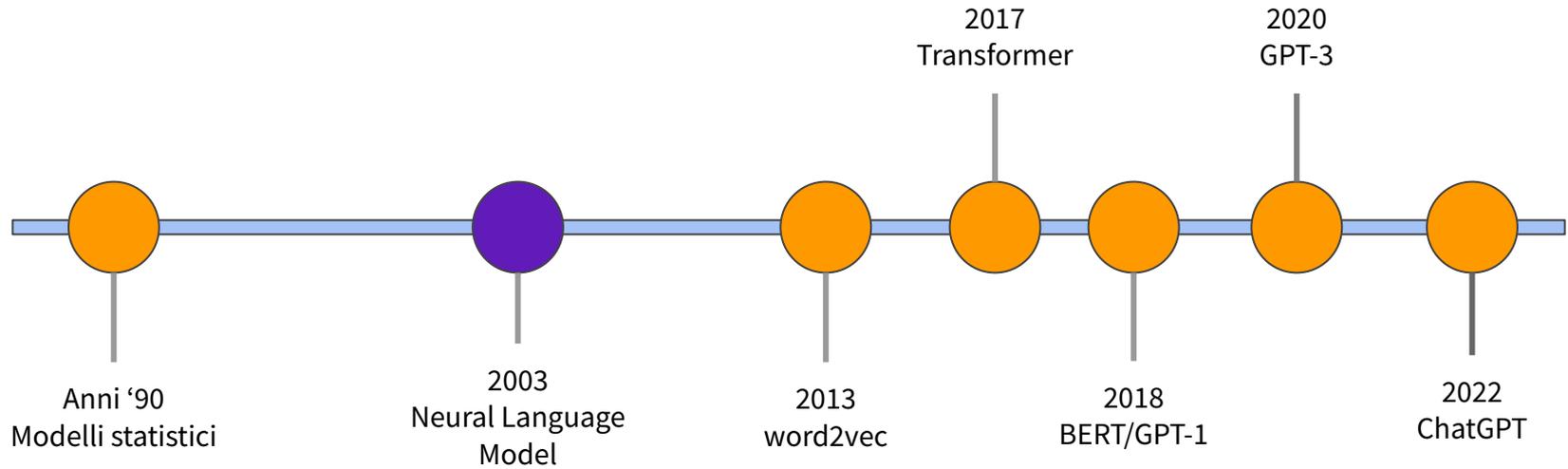
- Il settore dell’NLP ha visto un progresso senza precedenti nel corso degli ultimi anni
- Gran parte di questi progressi sono dovuti alla sostituzione dei “sistemi tradizionali” con più recenti modelli basati su algoritmi di Deep Learning



Timeline



Timeline



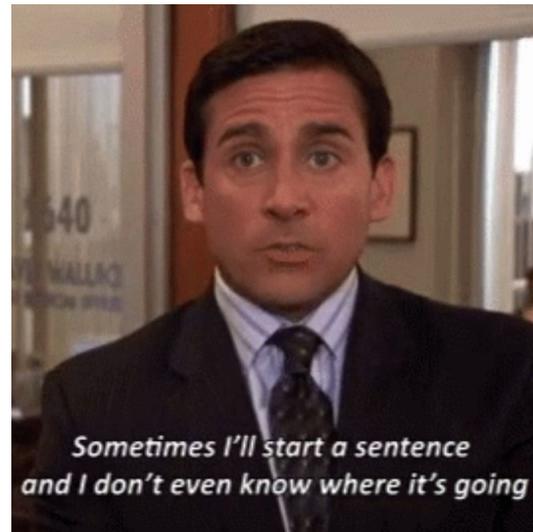
Neural Language Model (NLM)

- UN NLM è una rete neurale addestrata per approssimare la funzione di **language modeling**

Neural Language Model (NLM)

- UN NLM è una rete neurale addestrata per approssimare la funzione di **language modeling**
- Un modello del linguaggio probabilistico (**LM**) definisce la probabilità di una frase $s = [w_1, w_2, \dots, w_n]$ come:

$$P(s) = \prod_{i=1}^N P(w_i | w_1, w_2, \dots, w_{i-1})$$



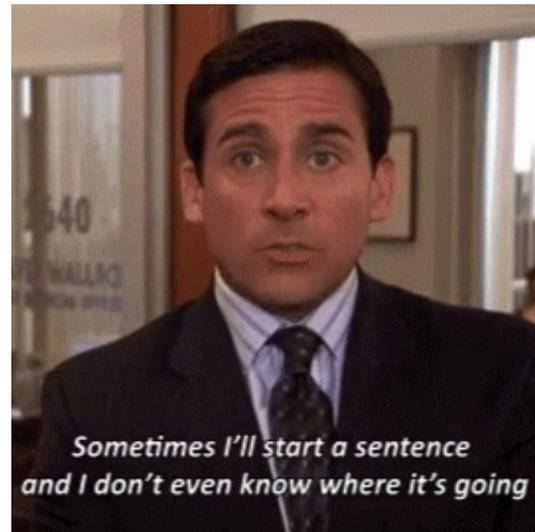
*Sometimes I'll start a sentence
and I don't even know where it's going*

Neural Language Model (NLM)

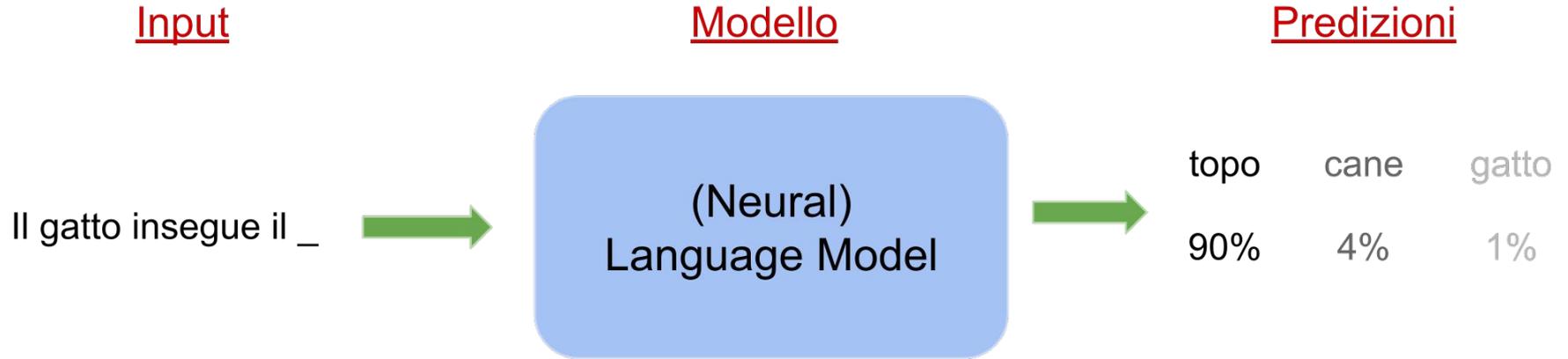
- UN NLM è una rete neurale addestrata per approssimare la funzione di **language modeling**
- Un modello del linguaggio probabilistico (**LM**) definisce la probabilità di una frase $s = [w_1, w_2, \dots, w_n]$ come:

$$P(s) = \prod_{i=1}^N P(w_i | w_1, w_2, \dots, w_{i-1})$$

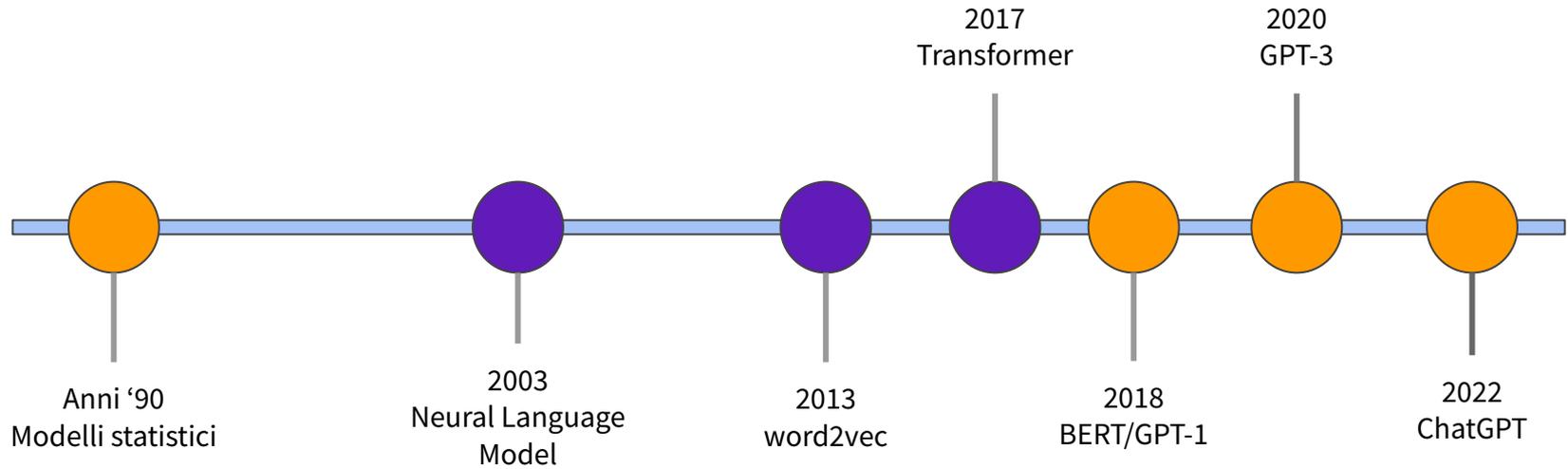
- **Bengio et al. (2003)** proposero un modello in grado di risolvere tale funzione ricorrendo all'architettura di una rete neurale → **Neural Probabilistic Language Model**



Neural Language Model (NLM)



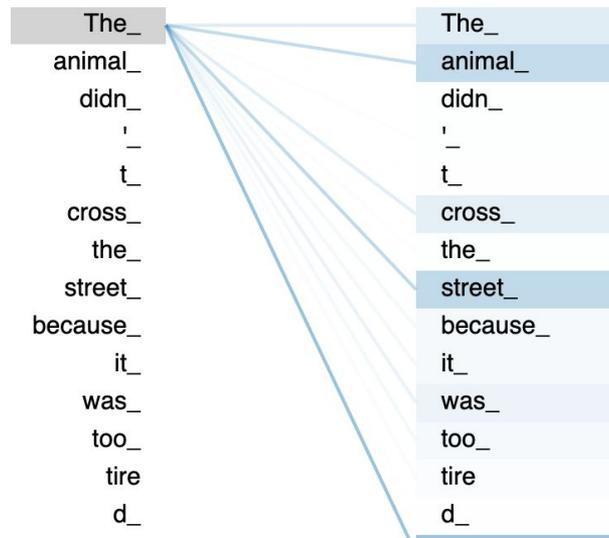
Timeline



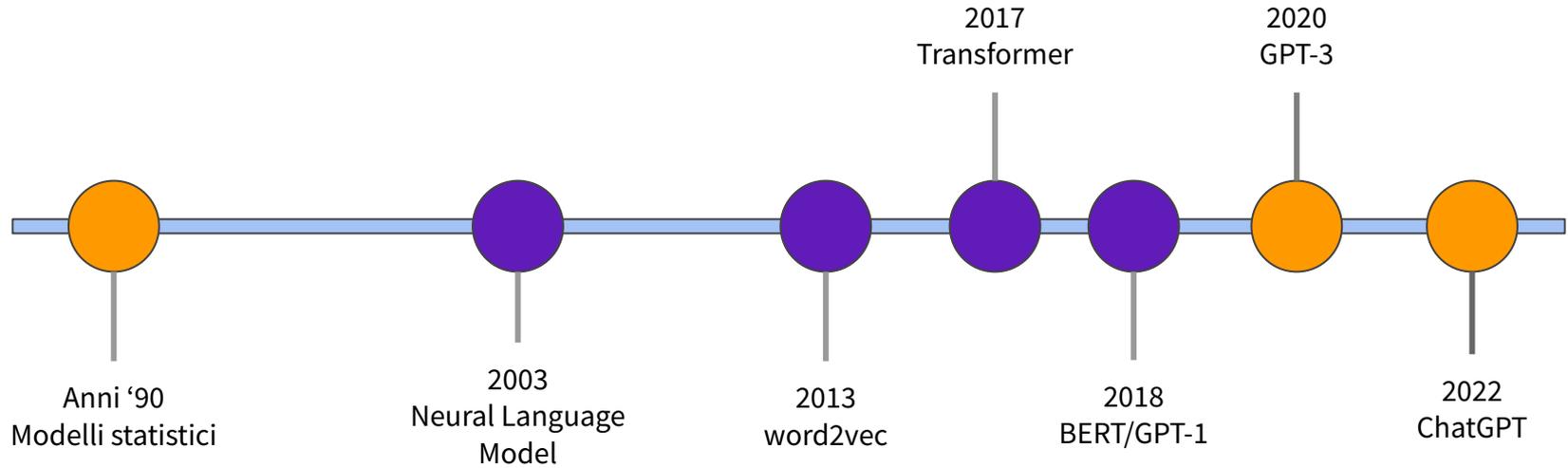
Transformer Models

- Ad oggi, il Transformer è l'architettura più utilizzata per lo sviluppo di NLM allo stato dell'arte
- Il Transformers (Vaswani et al., 2017) sfrutta il meccanismo dell'**attention** per creare rappresentazioni contestuali delle parole e imparare le relazioni fra di esse

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



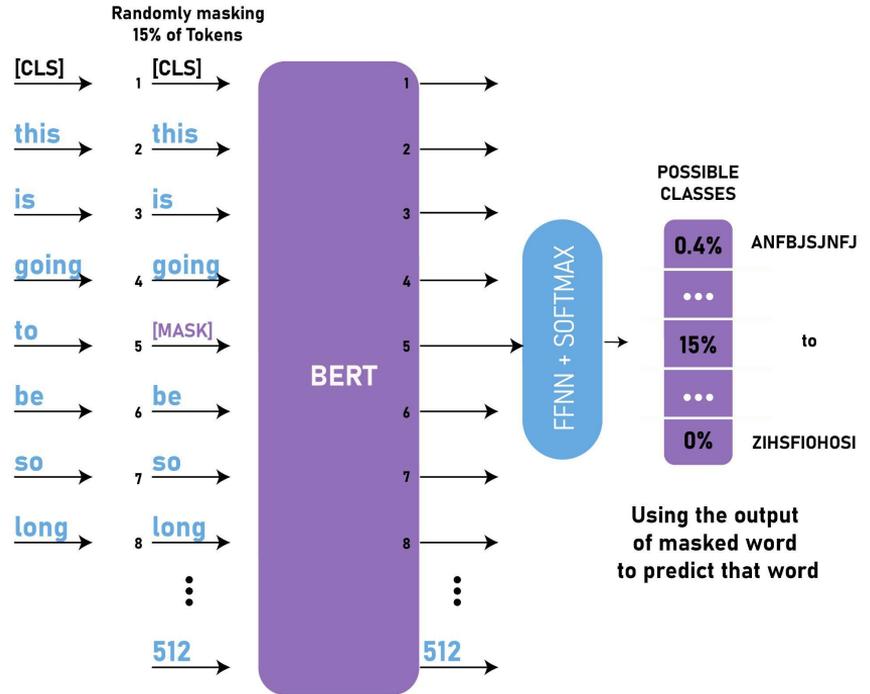
Timeline



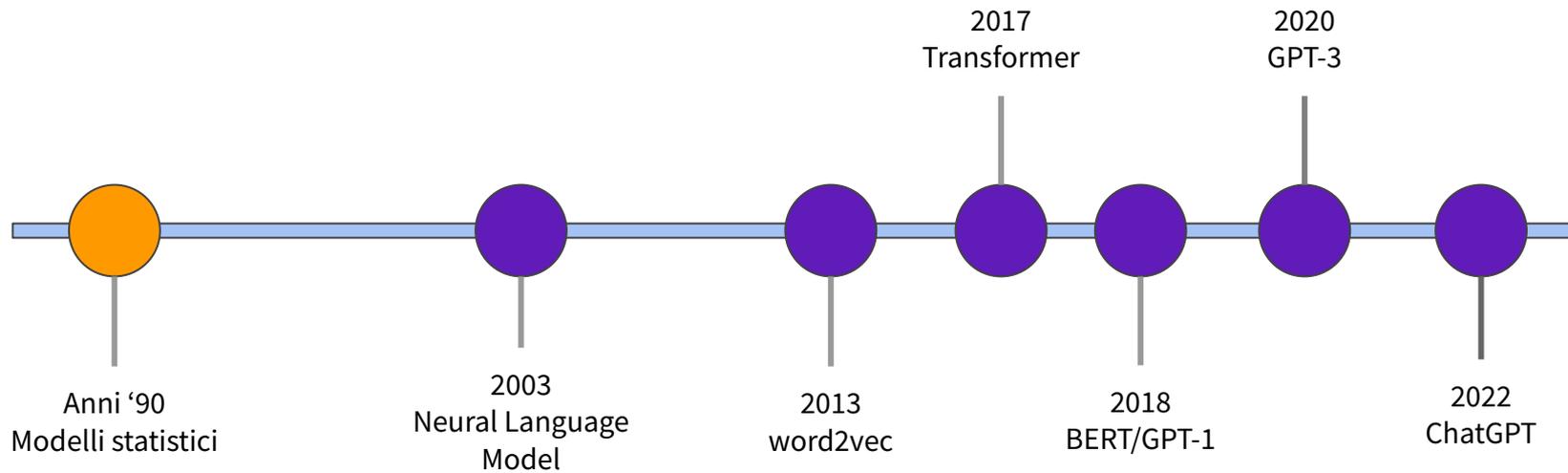
BERT (Devlin et al., 2019)



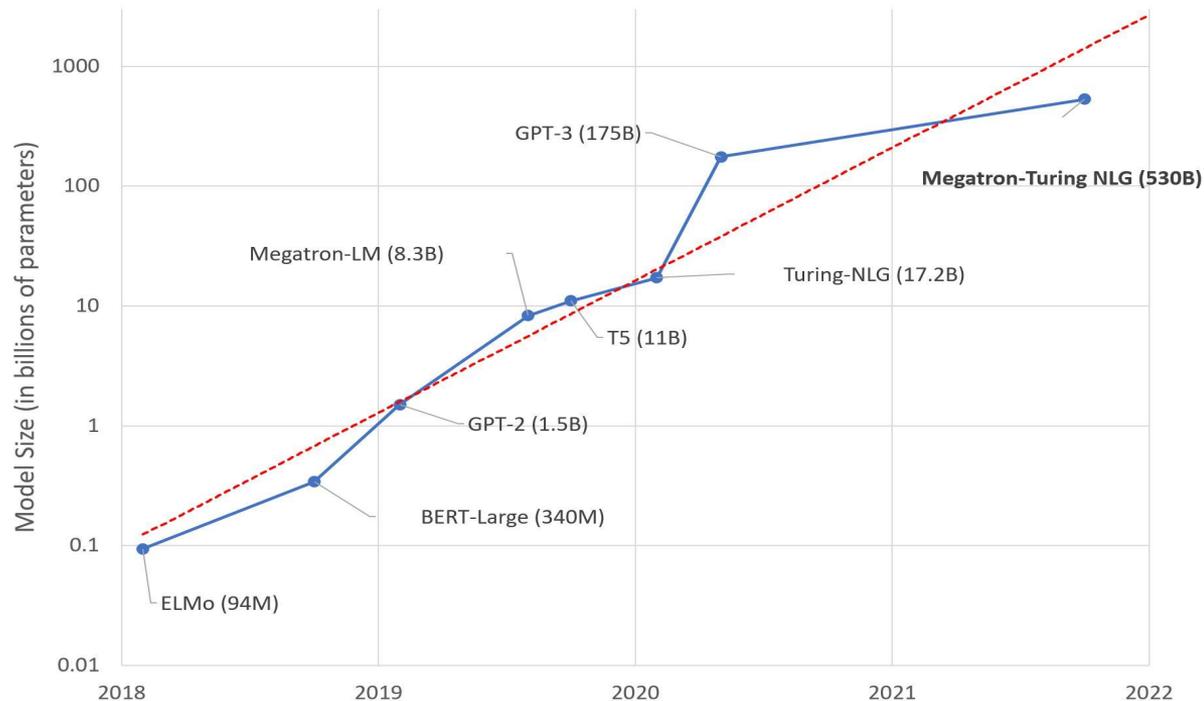
- Modello encoder (12/24 layers)
- Addestrato per approssimare la funzione di **Masked Language Modeling (MLM)**
- Il modello può poi essere ri-addestrato (fine-tuning) per risolvere svariati task di NLP:
 - Sentiment analysis;
 - Question answering;
 - Textual entailment;
 - etc.



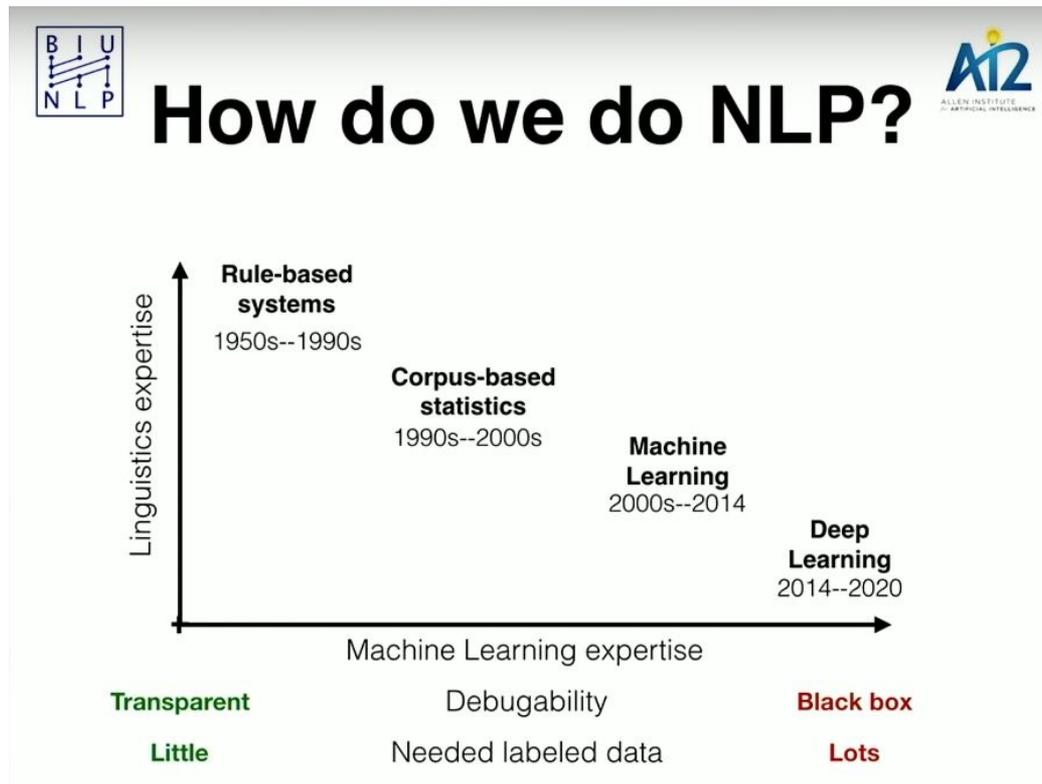
Timeline



Parameters are all you need (?)

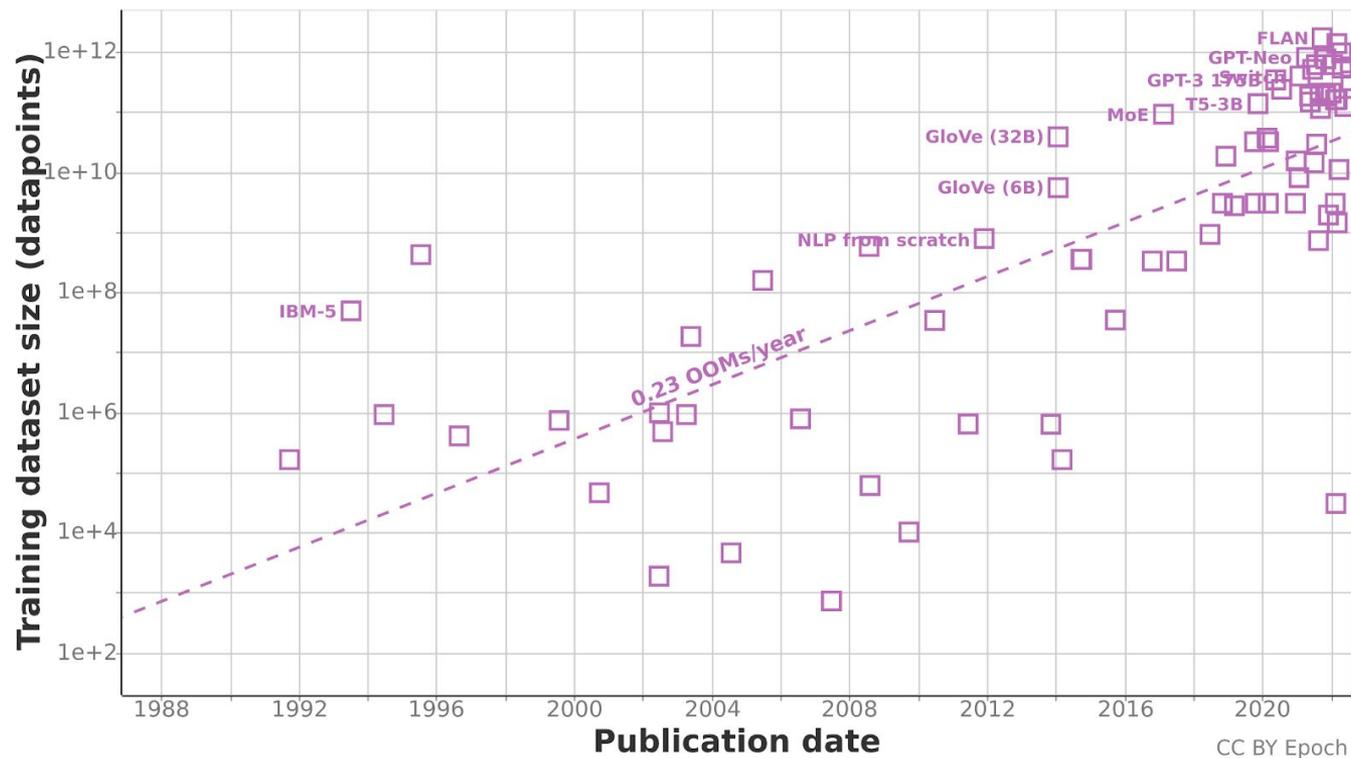


The missing elements in NLP (Yoav Goldberg, spaCy IRL 2019)



Il ruolo delle risorse linguistiche

Data is all you need



Data is all you need

Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion
Books1	12 billion
Books2	55 billion
Wikipedia	3 billion



Valutazione, specializzazione e interpretazione

Qual è il ruolo delle risorse linguistiche annotate nell'epoca dei Large Language Models addestrati su milioni di documenti contenenti testo grezzo?

Valutazione, specializzazione e interpretazione

Qual è il ruolo delle risorse linguistiche annotate nell'epoca dei Large Language Models addestrati su milioni di documenti contenenti testo grezzo?

- Il ruolo delle risorse linguistiche annotate non è meno importante, ma possiamo dire che è cambiato
- Principali utilizzi:
 - Valutazione dei modelli → Benchmark, campagne di valutazione;
 - Specializzazione dei modelli per task/competenze specifiche;
 - Analisi/interpretabilità dei modelli (probing tasks, diagnostic tests, ecc.).

Valutazione dei modelli

Valutazione dei modelli

AL

La frase "Carlo conosceva bene il compagno di classe che incontrare divertiva sempre Anna." è grammaticalmente corretta?



La frase "Carlo conosceva bene il compagno di classe che incontrare divertiva sempre Anna" è grammaticalmente corretta, ma può essere ambigua nella sua interpretazione a causa della posizione della frase "che incontrare". Questa frase può essere interpretata in due modi diversi:



1. Carlo conosceva bene il compagno di classe, il quale (che) incontrare divertiva sempre Anna.
In questa interpretazione, il compagno di classe era la persona che Anna trovava divertente incontrare.
2. Carlo conosceva bene il compagno di classe che Anna incontrava sempre con piacere.
In questa interpretazione, la frase "che incontrare" si riferisce all'azione di incontrare il compagno di classe, che Anna trovava divertente.

Per evitare l'ambiguità, si potrebbe riformulare la frase in modo più chiaro, ad esempio:
"Carlo conosceva bene il compagno di classe che Anna incontrava sempre con piacere e che la divertiva molto."

Valutazione dei modelli

- NLP benchmarks per la valutazione delle performance di (L)LMs;
- Alcuni esempi:
 - **GLUE Benchmark**
(<https://gluebenchmark.com/>):
 - Linguistic Acceptability;
 - Sentiment Analysis;
 - Textual Entailment;
 - etc
 - **SuperGLUE**
(<https://super.gluebenchmark.com/>):
 - Words in context;
 - Broadcoverage Diagnostics;
 - etc.



The General Language Understanding Evaluation (GLUE) benchmark is a collection of resources for training, evaluating, and analyzing natural language understanding systems. GLUE consists of:

- A benchmark of nine sentence- or sentence-pair language understanding tasks built on established existing datasets and selected to cover a diverse range of dataset sizes, text genres, and degrees of difficulty,
- A diagnostic dataset designed to evaluate and analyze model performance with respect to a wide range of linguistic phenomena found in natural language, and
- A public leaderboard for tracking performance on the benchmark and a dashboard for visualizing the performance of models on the diagnostic set.

Valutazione dei modelli



Dynabench: Rethinking Benchmarking in NLP
(Kiela et al., 2021),
<https://aclanthology.org/2021.naacl-main.324.pdf?ref=ruder-io>

Figure 1: Benchmark saturation over time for popular benchmarks, normalized with initial performance at minus one and human performance at zero.

Valutazione dei modelli

- E per la lingua italiana?
- **EVALITA** : campagna periodica (biennale) di valutazione dei modelli di NLP
 - **!!** Edizione 2023, 7-8 Settembre 2023, Parma **!!**



EVALITA
Evaluation of NLP and Speech Tools for Italian

Link al sito: <https://www.evalita.it/>

Affect

- **EMit** – Categorical Emotion Detection in Italian Social Media (O. Araque, S. Frenda, D. Nozza, V. Patti, R. Sprugnoli)
- **EmotivITA** – Dimensional and Multi-dimensional emotion analysis (G. Gafà, F. Cutugno, M. Venuti)

Authorship Analysis

- **Politicit** – Political Ideology Detection in Italian Texts (D. Russo, S.M. Jiménez-Zafra, J.A. García-Díaz, T. Caselli, M. Guerini, L.A. Ureña-López, R. Valencia-García)
- **GeoLangIt** – Geolocation of Linguistic Variation in Italy (A. Ramponi, C. Casula)
- **LangLearn** – Language Learning Development (C. Alzetta, D. Brunato, F. Dell'Orletta, A. Miaschi¹, K. Sagae, C.H. Sánchez-Gutiérrez, G. Venturi)

Computational Ethics

- **HaSpeeDe 3** – Political and Religious Hate Speech Detection (M. Lai, F. Celli, A. Ramponi, S. Tonelli, C. Bosco, V. Patti)
- **HODI** – Homotransphobia Detection in Italian (D. Nozza, G. Damo, A.T. Cignarella, T. Caselli, V. Patti)
- **MULTIFake-DetectIVE** – MULTImodal Fake News Detection and VERification (A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, L.C. Passaro)
- **ACTI** – Automatic Conspiracy Theory Identification (G. Russo, N. Stoehr, M. Horta Ribeiro)

New Challenges in Long-standing Tasks

- **NERMuD** – Named-Entities Recognition on Multi-Domain Documents (T. Paccosi, A. Palmero Arosio)
- **CLinkaRT** – Linking a Lab Result to its Test Event in the Clinical Domain (B. Magnini, B. Altuna, A. Lavelli, M. Speranza, R. Zanoli)
- **WIC-ITA** – Word-in-Context task for Italian (P. Cassotti, L. Siciliani, L. Passaro, M. Gatto, P. Basile)
- **DisCoTEX** – Assessing DIScourse COherence in Italian TEXTs (D. Brunato, D. Colla, F. Dell'Orletta, I. Dini, D.P. Radicioni, A.A. Ravelli)

Specializzazione dei modelli

Specializzazione dei modelli

T5 model

🗨️ Text2Text Generation

If you like original gut wrenching laughter you will like this movie. The sentiment of this sentence is <extra_id_0>.

Compute

ctrl+Enter

0.0

Computation time on Intel Xeon 3rd Gen Scalable cpu: 4.105 s

aresounding one.

Specializzazione dei modelli

T5 model

📄 Text2Text Generation

If you like original gut wrenching laughter you will like this movie. The sentiment of this sentence is <extra_id_0>.

Compute

ctrl+Enter

0.0

Computation time on Intel Xeon 3rd Gen Scalable cpu: 4.105 s

aresounding one.

T5 specializzato su taks di Sentiment Analysis

📄 Text2Text Generation

If you like original gut wrenching laughter you will like this movie. The sentiment of this sentence is <extra_id_0>.

Compute

ctrl+Enter

0.0

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.563 s

positive

Specializzazione dei modelli

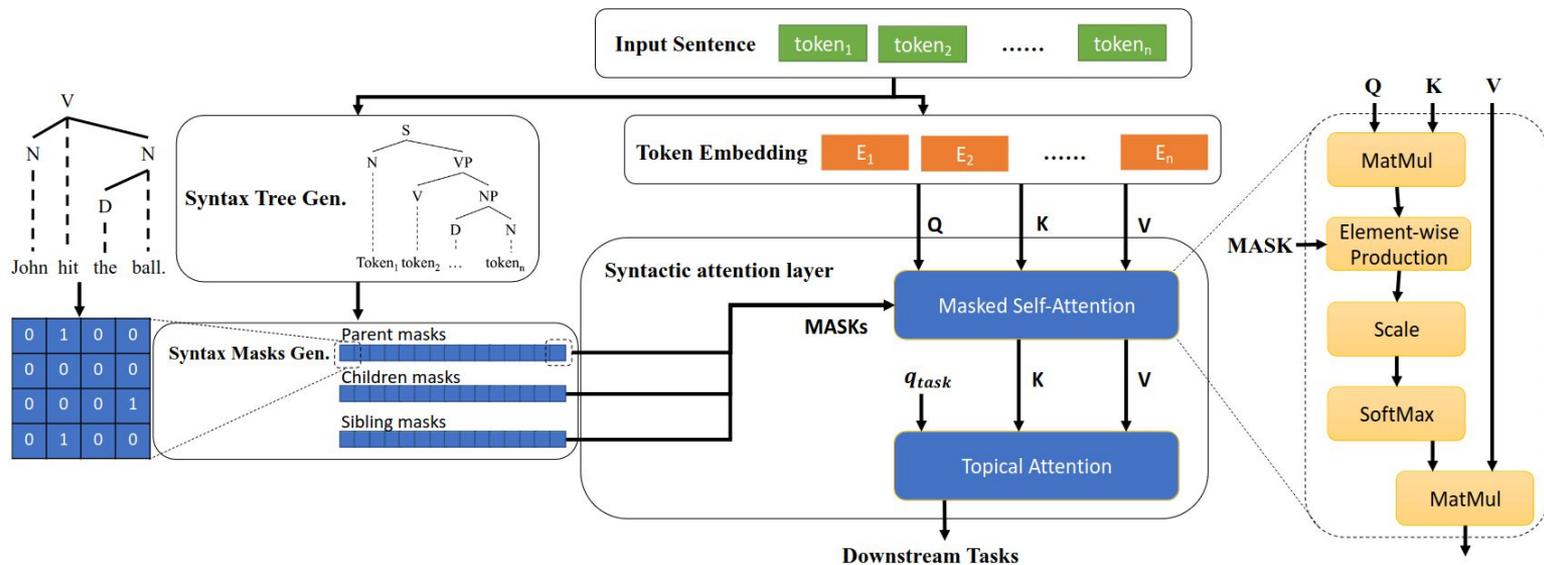


Figure 1: The Overall Architecture of Syntax-BERT. Note that the leftmost part shows an example of syntax tree and its corresponding parent syntax mask ($d = 1$).

Specializzazione dei modelli

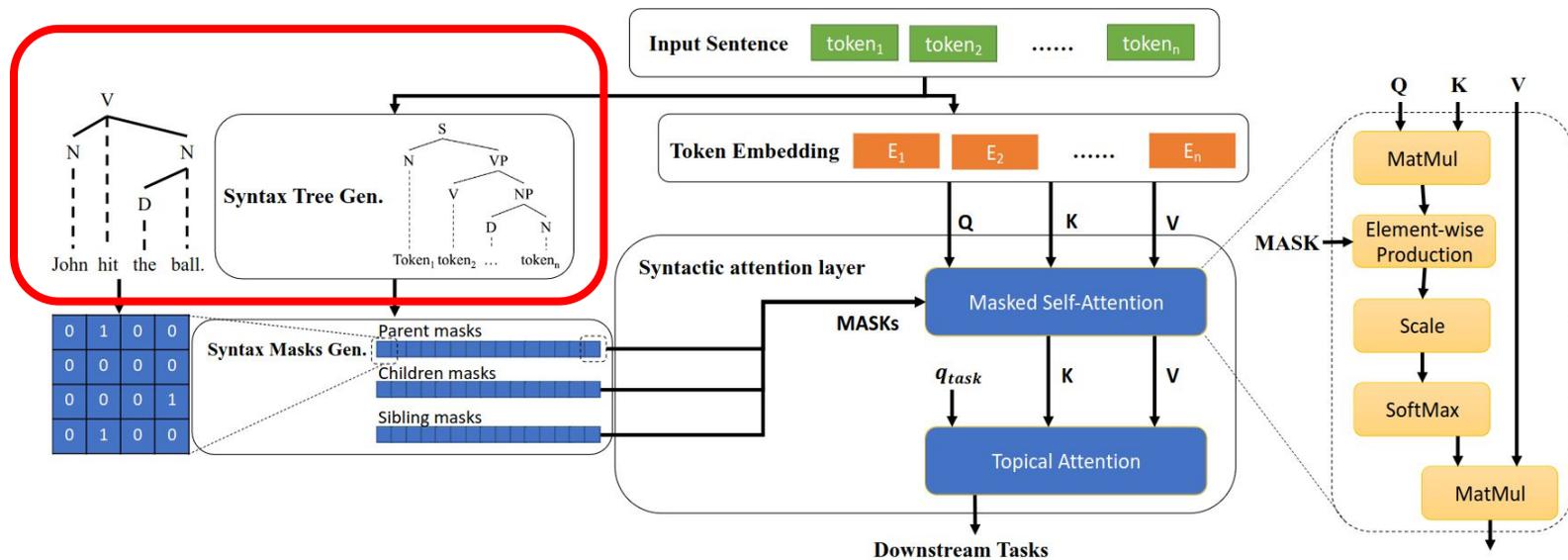


Figure 1: The Overall Architecture of Syntax-BERT. Note that the leftmost part shows an example of syntax tree and its corresponding parent syntax mask ($d = 1$).

Specializzazione dei modelli

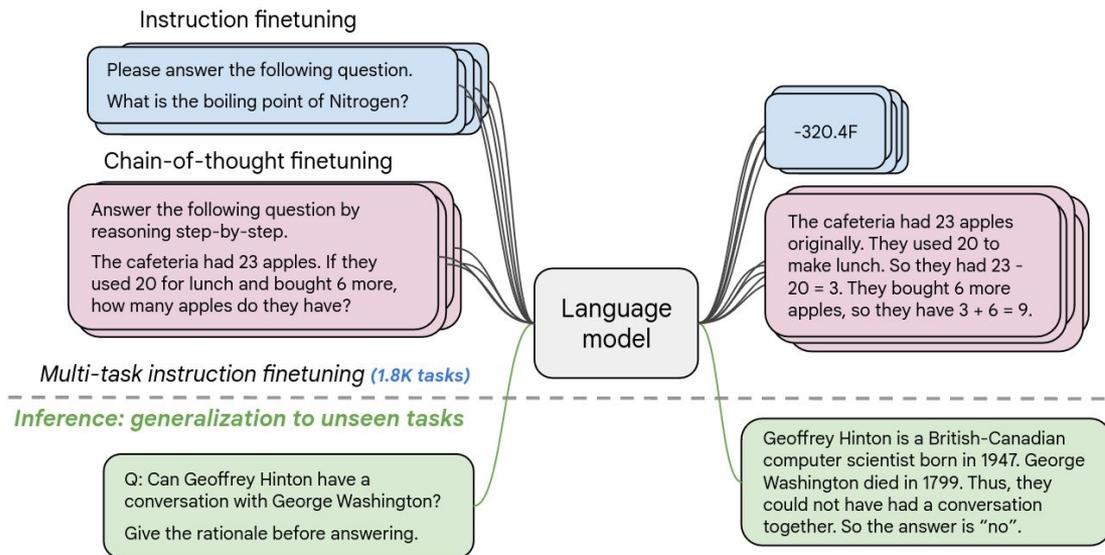
Model	SST-1	SST-2
Transformer	48.4	86.2
LISA-Transformer	52.2	89.1
Syntax-Transformer (Ours)	52.7	90.1
Syntax-Transformer + LISA (Ours)	53.2	91.1
BERT-Base	53.7	93.5
LISA-BERT-Base	54.2	93.7
Syntax-BERT-Base (Ours)	54.4	94.0
Syntax-BERT-Base + LISA (Ours)	54.5	94.4
BERT-Large	54.8	94.9
LISA-BERT-Large	55.0	95.9
Syntax-BERT-Large (Ours)	55.3	96.1
Syntax-BERT-Large + LISA (Ours)	55.5	96.4

Model	SNLI	MNLI
Transformer	84.9	71.4
LISA-Transformer	86.1	73.7
Syntax-Transformer (Ours)	86.8	74.1
Syntax-Transformer + LISA (Ours)	87.0	74.5
BERT-Base	87.0	84.3
LISA-BERT-base	87.4	84.7
Syntax-BERT-Base (Ours)	87.7	84.9
Syntax-BERT-Base + LISA (Ours)	87.8	84.9
BERT-Large	88.4	86.8
LISA-BERT-Large	88.8	86.8
Syntax-BERT-Large (Ours)	88.9	86.7
Syntax-BERT-Large + LISA (Ours)	89.0	87.0

Table 2: Comparison with SOTA models on SST dataset.

Table 3: Comparison with SOTA models on NLI datasets.

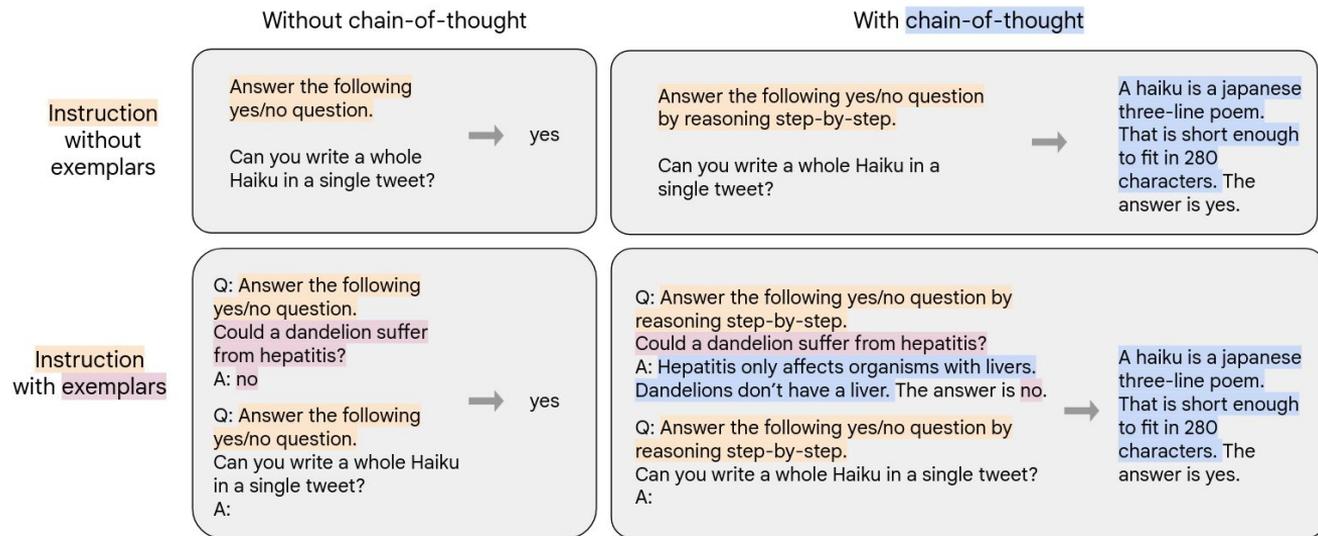
Specializzazione dei modelli



Scaling
Instruction-Finetuned
Language Models (Chung
et al., 2022),
<https://arxiv.org/pdf/2210.11416.pdf>

Figure 1: We finetune various language models on 1.8K tasks phrased as instructions, and evaluate them on unseen tasks. We finetune both with and without exemplars (i.e., zero-shot and few-shot) and with and without chain-of-thought, enabling generalization across a range of evaluation scenarios.

Specializzazione dei modelli



Scaling
Instruction-Finetuned
Language Models (Chung
et al., 2022),
<https://arxiv.org/pdf/2210.11416.pdf>

Figure 3: Combinations of finetuning data formats in this work. We finetune with and without exemplars, and also with and without chain-of-thought. In addition, we have some data formats without instructions but with few-shot exemplars only, like in Min et al. (2022) (not shown in the figure). Note that only nine chain-of-thought (CoT) datasets use the CoT formats.

Specializzazione dei modelli

Model	Finetuning Mixtures	Tasks	Norm. avg.	MMLU		BBH		TyDiQA	MGSM
				Direct	CoT	Direct	CoT	Direct	CoT
8B	None (no finetuning)	0	6.4	24.3	24.1	30.8	30.1	25.0	3.4
	CoT	9	8.3 (+1.9)	26.3	32.1	19.8	26.6	39.3	<u>10.4</u>
	CoT, Muffin	89	14.8 (+8.4)	37.6	38.4	31.0	30.9	32.4	8.4
	CoT, Muffin, T0-SF	282	20.5 (+14.1)	47.7	39.7	33.1	30.9	<u>49.0</u>	8.5
	CoT, Muffin, T0-SF, NIV2	1,836	<u>21.9</u> (+15.5)	<u>49.3</u>	<u>41.3</u>	<u>36.4</u>	<u>31.1</u>	47.5	8.2
62B	None (no finetuning)	0	28.4	55.1	49.0	37.4	43.0	40.5	18.2
	CoT	9	29.0 (+0.4)	48.5	48.7	34.5	39.5	48.8	<u>32.6</u>
	CoT, Muffin	89	33.4 (+6.0)	55.3	51.4	42.8	40.2	53.0	23.9
	CoT, Muffin, T0-SF	282	37.9 (+9.5)	<u>60.0</u>	56.0	44.7	43.8	58.2	30.0
	CoT, Muffin, T0-SF, NIV2	1,836	<u>38.8</u> (+10.4)	59.6	<u>56.9</u>	<u>47.5</u>	<u>44.9</u>	<u>58.7</u>	28.5
540B	None (no finetuning)	0	49.1	71.3	62.9	49.1	63.7	52.9	45.9
	CoT	9	52.6 (+3.5)	68.8	64.8	50.5	61.1	61.2	59.4
	CoT, Muffin	89	57.0 (+7.9)	71.8	66.7	56.7	64.0	65.3	<u>63.0</u>
	CoT, Muffin, T0-SF	282	57.5 (+8.4)	72.9	<u>68.2</u>	57.3	64.0	65.8	61.6
	CoT, Muffin, T0-SF, NIV2	1,836	<u>58.5</u> (+9.4)	<u>73.2</u>	68.1	<u>58.8</u>	<u>65.6</u>	<u>67.4</u>	61.3

Table 3: Increasing the number of tasks in the finetuning data improves performance of Flan-PaLM on most evaluation benchmarks. The benchmark suites are MMLU (57 tasks), BBH (23 tasks), TyDiQA (8 languages), and MGSM (10 languages). The evaluation metric on all four benchmark suites is few-shot prompted accuracy (exact match), where we take an unweighted average over all tasks. As an aggregate metric we report the normalized average of MMLU-direct, MMLU-CoT, BBH-direct, BBH-CoT, TyDiQA, and MGSM. These evaluation benchmarks are held-out (not included in the finetuning data).

Scaling
Instruction-Finetuned
Language Models (Chung
et al., 2022),
<https://arxiv.org/pdf/2210.11416.pdf>

Specializzazione dei modelli

Model	Finetuning Mixtures	Tasks	Norm. avg.	MMLU		BBH		TyDiQA	MGSM
				Direct	CoT	Direct	CoT	Direct	CoT
8B	None (no finetuning)	0	6.4	24.3	24.1	30.8	30.1	25.0	3.4
	CoT	9	8.3 (+1.9)	26.3	32.1	19.8	26.6	39.3	10.4
	CoT, Muffin	89	14.8 (+8.4)	37.6	38.4	31.0	30.9	32.4	8.4
	CoT, Muffin, T0-SF	282	20.5 (+14.1)	47.7	39.7	33.1	30.9	49.0	8.5
	CoT, Muffin, T0-SF, NIV2	1,836	21.9 (+15.5)	49.3	41.3	36.4	31.1	47.5	8.2
62B	None (no finetuning)	0	28.4	55.1	49.0	37.4	43.0	40.5	18.2
	CoT	9	29.0 (+0.4)	48.5	48.7	34.5	39.5	48.8	32.6
	CoT, Muffin	89	33.4 (+6.0)	55.3	51.4	42.8	40.2	53.0	23.9
	CoT, Muffin, T0-SF	282	37.9 (+9.5)	60.0	56.0	44.7	43.8	58.2	30.0
	CoT, Muffin, T0-SF, NIV2	1,836	38.8 (+10.4)	59.6	56.9	47.5	44.9	58.7	28.5
540B	None (no finetuning)	0	49.1	71.3	62.9	49.1	63.7	52.9	45.9
	CoT	9	52.6 (+3.5)	68.8	64.8	50.5	61.1	61.2	59.4
	CoT, Muffin	89	57.0 (+7.9)	71.8	66.7	56.7	64.0	65.3	63.0
	CoT, Muffin, T0-SF	282	57.5 (+8.4)	72.9	68.2	57.3	64.0	65.8	61.6
	CoT, Muffin, T0-SF, NIV2	1,836	58.5 (+9.4)	73.2	68.1	58.8	65.6	67.4	61.3

Table 3: Increasing the number of tasks in the finetuning data improves performance of Flan-PaLM on most evaluation benchmarks. The benchmark suites are MMLU (57 tasks), BBH (23 tasks), TyDiQA (8 languages), and MGSM (10 languages). The evaluation metric on all four benchmark suites is few-shot prompted accuracy (exact match), where we take an unweighted average over all tasks. As an aggregate metric we report the normalized average of MMLU-direct, MMLU-CoT, BBH-direct, BBH-CoT, TyDiQA, and MGSM. These evaluation benchmarks are held-out (not included in the finetuning data).

Scaling
Instruction-Finetuned
Language Models (Chung
et al., 2022),
<https://arxiv.org/pdf/2210.11416.pdf>

Interpretazione dei modelli

Interpretazione dei modelli

- Nel contesto degli studi di NLP, interpretare un NLM vuol dire definire tecniche per:
 - comprendere le competenze linguistiche codificate implicitamente da tali modelli;
 - studiare come tali competenze sono (o non sono) utilizzate nel momento in cui un dato modello viene addestrato per risolvere un preciso *downstream/fine-tuning task*

Interpretazione dei modelli

- Nel contesto degli studi di NLP, interpretare un NLM vuol dire definire tecniche per:
 - comprendere le competenze linguistiche codificate implicitamente da tali modelli;
 - studiare come tali competenze sono (o non sono) utilizzate nel momento in cui un dato modello viene addestrato per risolvere un preciso *downstream/fine-tuning task*

- Alcune domande:
 - Quali competenze sintattiche sono codificate all'interno delle rappresentazioni interne (word/sentence embeddings) di un NLM?
 - Quali competenze linguistiche vengono utilizzate quando il modello risolve un task, e.g. Sentiment Analysis?

Assessing BERT's Syntactic Abilities (Goldberg, 2019)

- Goldberg (2019) propone una metodologia per testare la competenza linguistica implicita di BERT
- Nello specifico, sono considerati due fenomeni specifici:
 - Accordo Soggetto-Verbo;
 - Anafora.
- **Approccio:** mascherare la parola target e chiedere al modello di completare con la parola predetta con probabilità più alta

Assessing BERT's Syntactic Abilities (Goldberg, 2019)

the game that the guard hates is bad

Assessing BERT's Syntactic Abilities (Goldberg, 2019)

the game that the guard hates [**MASK**] bad

Assessing BERT's Syntactic Abilities (Goldberg, 2019)

the game that the guard hates **[MASK]** bad

- 
- $p(is) = ?$
 - $p(are) = ?$

Assessing BERT's Syntactic Abilities (Goldberg, 2019)

	BERT Base	BERT Large	LSTM (M&L)	Humans (M&L)	# Pairs (# M&L Pairs)
SUBJECT-VERB AGREEMENT:					
Simple	1.00	1.00	0.94	0.96	120 (140)
In a sentential complement	0.83	0.86	0.99	0.93	1440 (1680)
Short VP coordination	0.89	0.86	0.90	0.82	720 (840)
Long VP coordination	0.98	0.97	0.61	0.82	400 (400)
Across a prepositional phrase	0.85	0.85	0.57	0.85	19440 (22400)
Across a subject relative clause	0.84	0.85	0.56	0.88	9600 (11200)
Across an object relative clause	0.89	0.85	0.50	0.85	19680 (22400)
Across an object relative (no <i>that</i>)	0.86	0.81	0.52	0.82	19680 (22400)
In an object relative clause	0.95	0.99	0.84	0.78	15960 (22400)
In an object relative (no <i>that</i>)	0.79	0.82	0.71	0.79	15960 (22400)
REFLEXIVE ANAPHORA:					
Simple	0.94	0.92	0.83	0.96	280 (280)
In a sentential complement	0.89	0.86	0.86	0.91	3360 (3360)
Across a relative clause	0.80	0.76	0.55	0.87	22400 (22400)

Table 3: Results on the [Marvin and Linzen \(2018\)](#) stimuli. M&L results numbers are taken from [Marvin and Linzen \(2018\)](#). The BERT and M&L numbers are *not* directly comparable, as the experimental setup differs in many ways.

SyntaxGym (Gauthier et al., 2020)



SyntaxGym

SyntaxGym is a unified platform for targeted syntactic evaluation of language models. The Gym supports all steps of the evaluation process, from designing test suites to visualizing final results. Our goal is to make psycholinguistic assessment of language models more **standardized**, **reproducible**, and **accessible** to a wide variety of researchers.

TEST SUITES

Create new psycholinguistic test suites, or browse existing ones in our database.



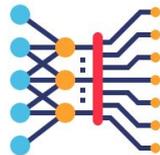
39 available suites

See more →



LANGUAGE MODELS

Evaluate a set of neural language models ranging in architecture and size.



9 available models

See more →



VISUALIZATIONS

Visualize results across models and test suites through interactive charts.



See more →



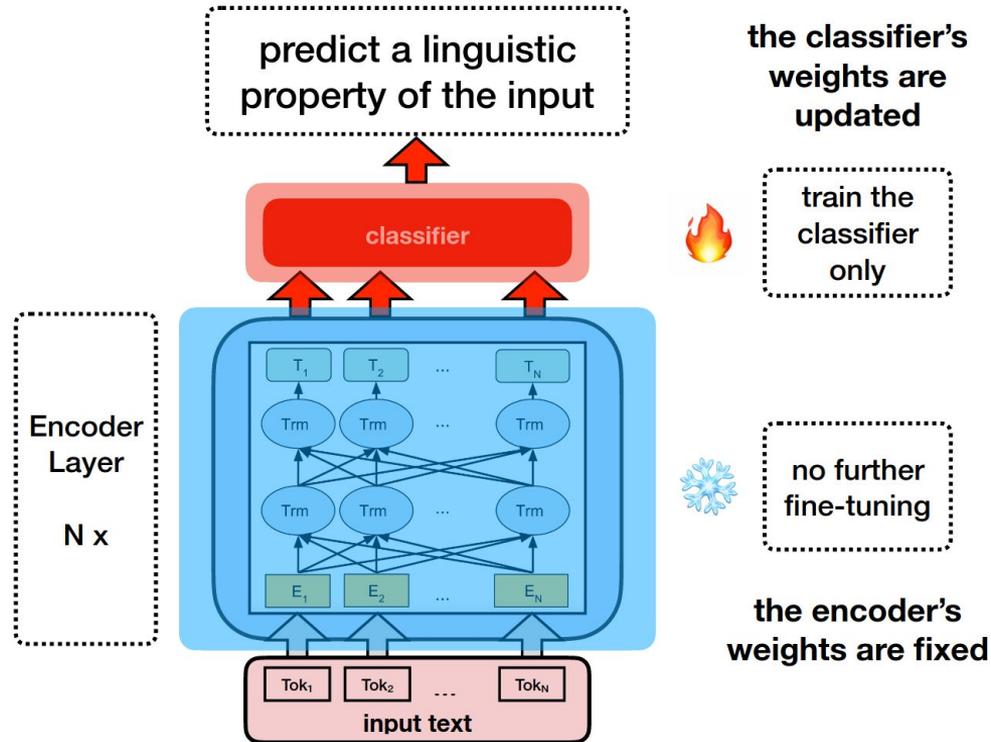
SyntaxGym: An Online Platform For Targeted Evaluation of Language Models (Gauthier et al., 2020)

Paper:

<https://aclanthology.org/2020.acl-demos.10/>

Sito: <https://syntaxgym.org/>

Probing Task



Profilare un Neural Language Model

- Il “*linguistic profiling*” (van Halteren, 2004) è una metodologia che assume che grandi quantità di caratteristiche (features) linguistiche possano essere particolarmente utili per la risoluzione di diversi task di NLP, e.g.:
 - Text Profiling (e.g. leggibilità, classificazione di genere testuale)
 - Author Profiling (e.g. età di uno scrittore; lingua materna di uno studente)

Domanda di ricerca:

È possibile sfruttare il potenziale di tali caratteristiche per studiare il comportamento implicito di NLMs allo stato dell'arte?

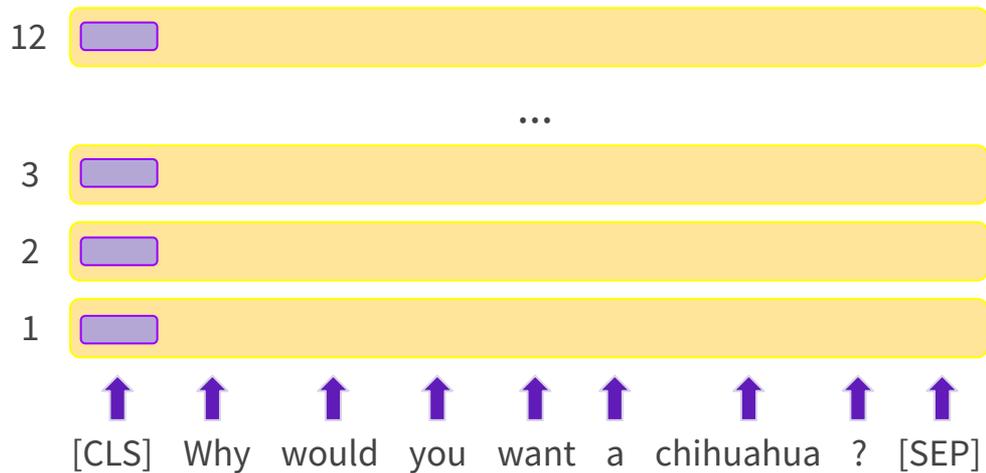
Linguistic Profiling of a Neural Language Model (Miaschi et al., 2020) [Outstanding Paper for COLING 2020],
<https://aclanthology.org/2020.coling-main.65.pdf>

Profiling-UD: a tool for Linguistic Profiling of Texts

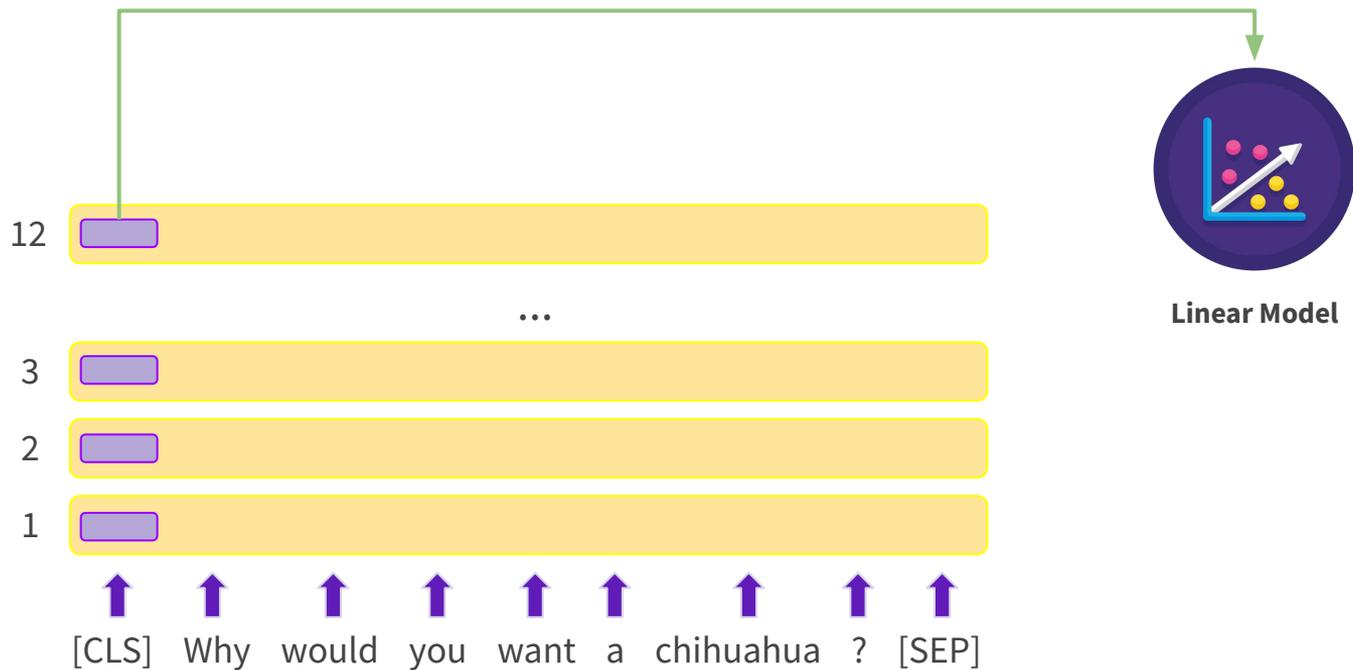
- ProfilingUD (Brunato et al., 2020) è un'applicazione che permette di effettuare il *linguistic profiling* di collezioni di testi per molteplici lingue
- Permette l'estrazione di più di 130 features, riconducibili a diversi livelli di annotazione linguistica
- Link: <http://linguistic-profiling.italianlp.it/>

Linguistic Feature
Raw Text Properties
Sentence Length
Word Length
Vocabulary Richness
Type/Token Ratio for words and lemmas
Morphosyntactic information
Distribution of UD and language-specific POS
Lexical density
Inflectional morphology
Inflectional morphology of lexical verbs and auxiliaries
Verbal Predicate Structure
Distribution of verbal heads and verbal roots
Verb arity and distribution of verbs by arity
Global and Local Parsed Tree Structures
Depth of the whole syntactic tree
Average length of dependency links and of the longest link
Average length of prepositional chains and distribution by depth
Clause length
Relative order of elements
Order of subject and object
Syntactic Relations
Distribution of dependency relations
Use of Subordination
Distribution of subordinate and principal clauses
Average length of subordination chains and distribution by depth
Relative order of subordinate clauses

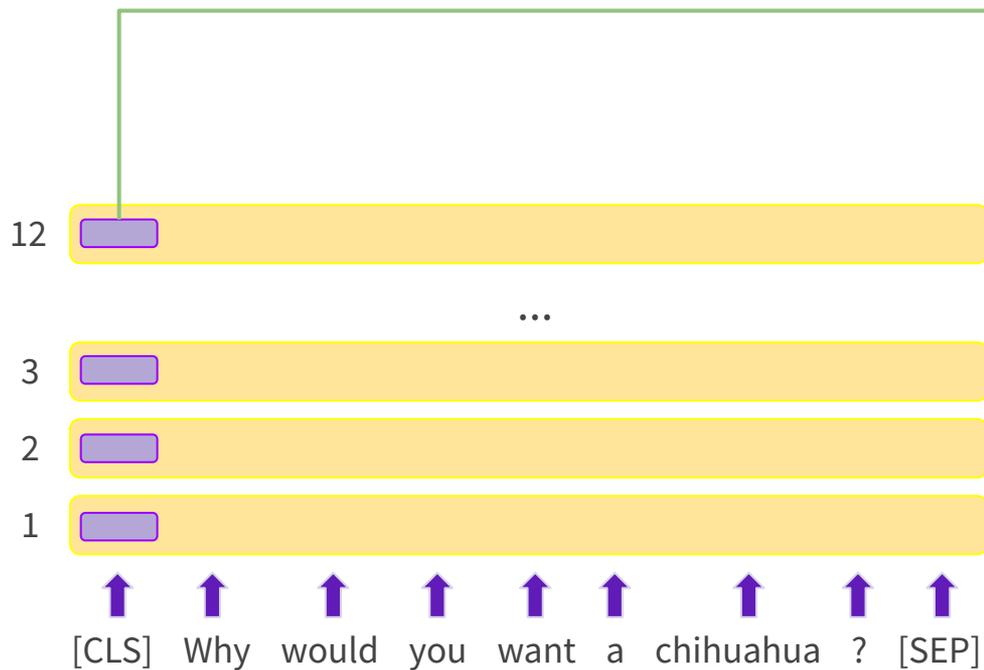
Profilare un Neural Language Model



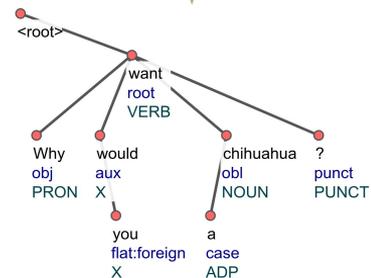
Profilare un Neural Language Model



Profilare un Neural Language Model



Linear Model



Profilare un Neural Language Model

- Che tipo di competenze linguistiche hanno i NLM italiani?
- In “*Probing Linguistic Knowledge in Italian Neural Language Models across Language Varieties*” (Miaschi et al. 2022), abbiamo applicato il linguistic profiling approach sui 7 Transformer model disponibili per la lingua italiana, al fine di:
 - Compararne le performance;
 - Indagare se e come la conoscenza codificata da questi NLM differisce tra generi testuali e le varietà linguistiche.



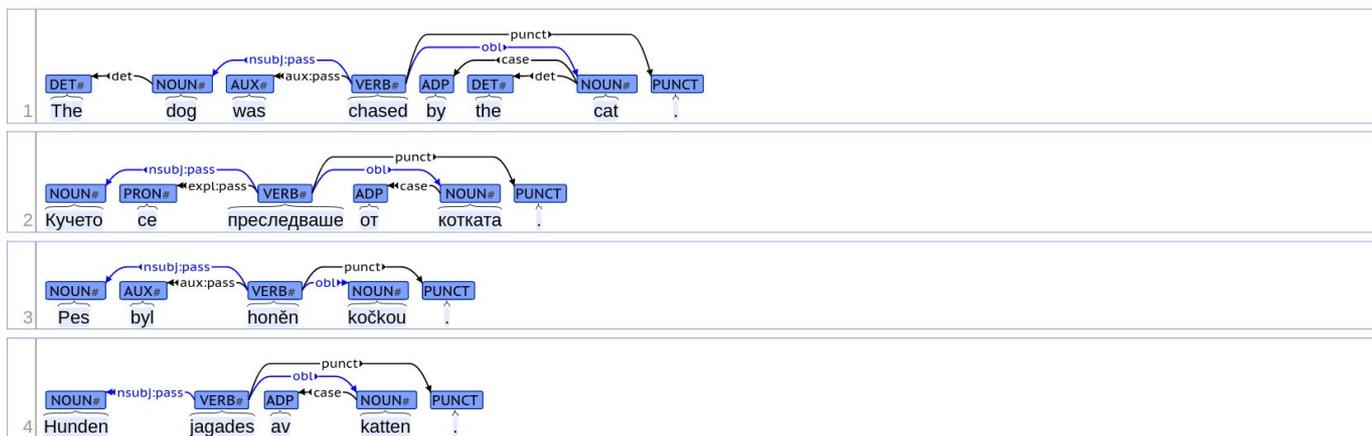
I modelli

Name	Training data
BERT Architecture	
Multilingual-BERT	Wikipedia
BERT-base-italian	Wikipedia + OPUS (13GB)
ALBERTo	TWITA (191GB)
RoBERTa Architecture	
GilBERTo	OSCAR (71GB)
UmBERTo-Commoncrawl	OSCAR (69GB)
UmBERTo-Wikipedia	Wikipedia (7GB)
GPT-2 Architecture	
GePpeTto	Wikipedia + ItWAC (14GB)

La risorsa linguistica

Universal Dependencies

Universal Dependencies (UD) is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages. UD is an open community effort with over 300 contributors producing nearly 200 treebanks in over 100 languages. If you're new to UD, you should start by reading the first part of the Short Introduction and then browsing the annotation guidelines.

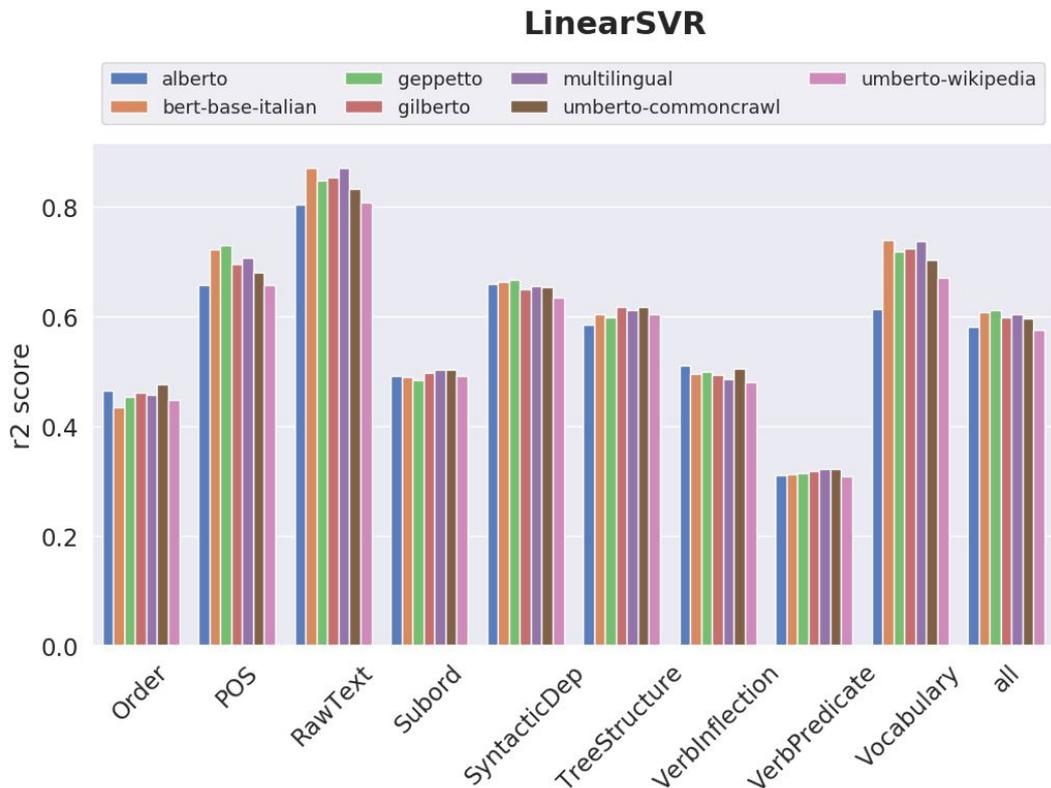


Link al sito: <https://universaldependencies.org/>

La risorsa linguistica

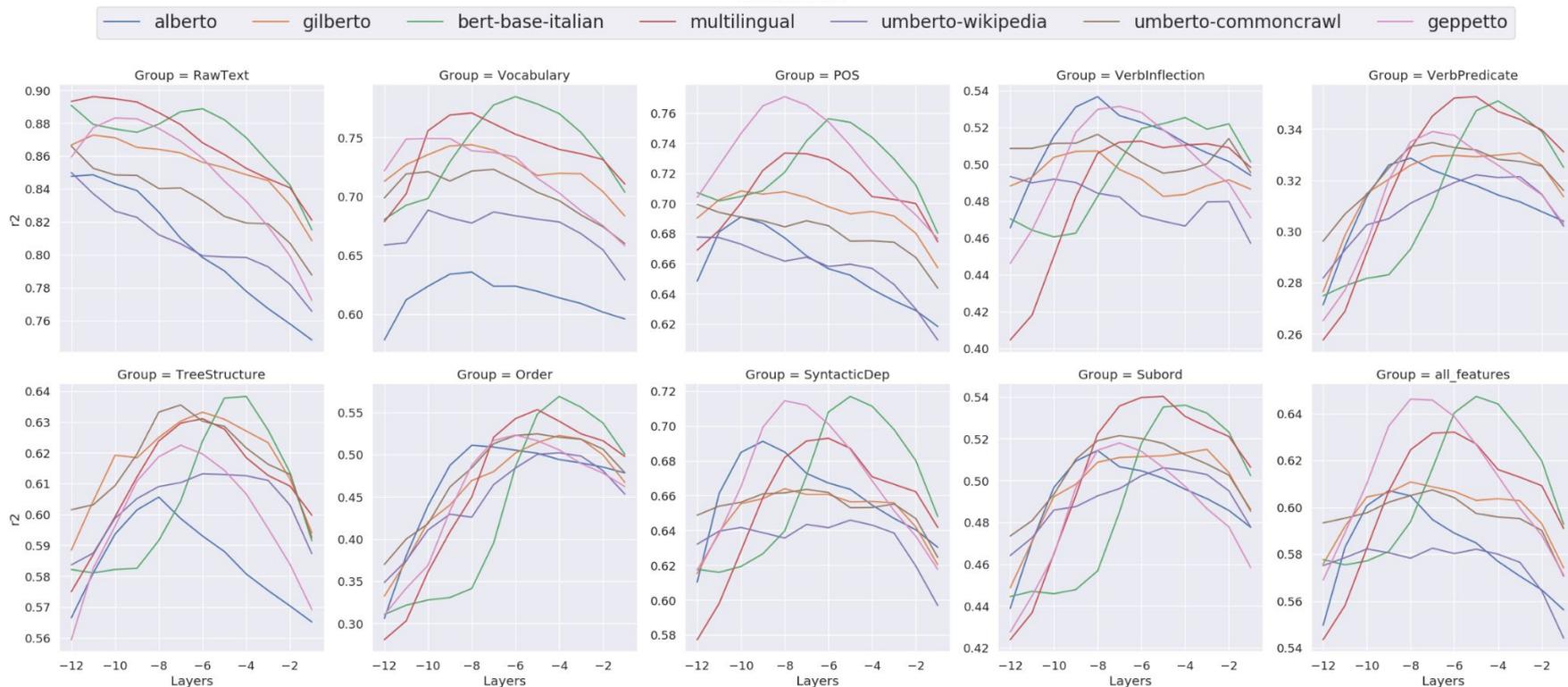
Short Name	Types of texts	# sent
ParTUT	Multi-genre	2,090
VIT	Multi-genre	10,087
ISDT	Multi-genre	14,167
ISDT_tanl	News wire	4,043
ISDT_tut	Legal/News wire/Wiki	3,802
ISDT_quest	Interrogative sentences	2,162
ISDT_2parole	Simplified Italian news	1,421
ISDT_europarl	EU Parliament debates	497
PoSTWITA	Tweets	6,713
TWITTIRÒ	Ironic Tweets	1,424
Total		35,481

I risultati

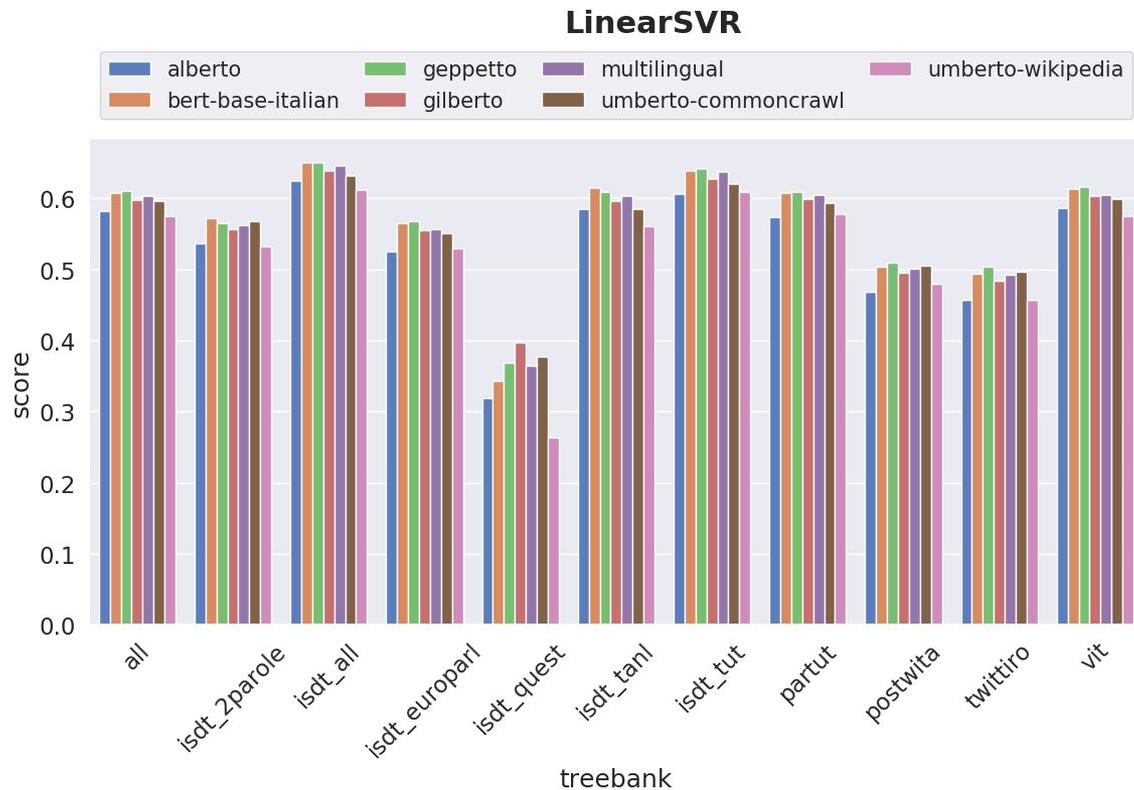


I risultati

LinearSVR



I risultati



AL

Credi che le risorse linguistiche annotate abbiano ancora un ruolo oggi, nell'epoca dei modelli del linguaggio come ChatGPT?



Absolutamente sì, le risorse linguistiche annotate hanno ancora un ruolo cruciale nell'epoca dei modelli di linguaggio come ChatGPT.



Mentre i modelli di linguaggio come ChatGPT sono in grado di apprendere da grandi quantità di dati non annotati, le risorse linguistiche annotate sono ancora necessarie per:

1. **Addestramento dei modelli:** le risorse linguistiche annotate sono spesso utilizzate come dati di addestramento per le parti specifiche del modello. Ad esempio, un modello di traduzione automatica potrebbe essere addestrato utilizzando un corpus parallelo, in cui ogni frase in una lingua è accompagnata da una traduzione nella lingua di destinazione.
2. **Valutazione dei modelli:** le risorse linguistiche annotate vengono utilizzate per valutare la qualità dei modelli di linguaggio. Ad esempio, un modello di classificazione del testo può essere valutato utilizzando un insieme di dati di test annotati manualmente, in cui ogni testo è associato a una o più etichette.
3. **Sviluppo di nuove applicazioni:** le risorse linguistiche annotate sono spesso utilizzate come base per lo sviluppo di nuove applicazioni di elaborazione del linguaggio naturale. Ad esempio, un sistema di riconoscimento vocale potrebbe essere addestrato utilizzando un corpus di registrazioni audio annotate manualmente.

Grazie per l'attenzione!



<https://alemiaschi.github.io/>



[@AlessioMiaschi](https://twitter.com/AlessioMiaschi)



<http://www.italianlp.it/>



[@ItaliaNLP_Lab](https://twitter.com/ItaliaNLP_Lab)

Riferimenti

- Bengio, Yoshua, et al (2003). "A neural probabilistic language model." *The journal of machine learning research* 3: 1137-1155.
- Vaswani, Ashish, et al. (2017). "Attention is all you need". In *Advances in Neural Information Processing Systems* (NEURIPS), 30.
- Devlin, Jacob, et al. (2019) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.
- Brown, Tom, et al. (2020). "Language models are few-shot learners". In *Advances in Neural Information Processing Systems* (NEURIPS), 33, 1877-1901.
- Kiela, Douwe, et al. (2021), "Dynabench: Rethinking Benchmarking in NLP". In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL), pp. 4110-4124.
- Bai, Jiangang, et al. (2021). "Syntax-BERT: Improving Pre-trained Transformers with Syntax Trees". In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (EACL), pp. 3011-3020.
- Chung, Hyung, et al. (2022). "Scaling Instruction-finetuned Language Models". arXiv preprint arXiv:2210.11416.
- Gauthier, Jon, et al. (2020). "SyntaxGym: An Online Platform for Targeted Evaluation of Language Models". In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (ACL), pp. 70-76.
- Goldberg, Yoav. "Assessing BERT's syntactic abilities." *arXiv preprint arXiv:1901.05287* (2019).
- Miaschi, Alessio, et al. (2020). "Linguistic Profiling of a Neural Language Model". In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online).
- Miaschi, Alessio, et al. (2022). "Probing Linguistic Knowledge in Italian Neural Language Models across Language Varieties." *IJCoL. Italian Journal of Computational Linguistics* 8.8-1.