



How shall a machine call a thing?

Exploring Basicness in Language through Attention-based Neural Networks and Human-in-the-Loop Methodology



Federico Torrielli, 14/04/2023

Relatore: *Prof. Luigi Di Caro*
Contro-relatore: *Prof. Valerio Basile*



Introduzione

Il linguaggio è la **proprietà evolutiva** che primariamente ci differenzia da qualsiasi altro animale.

Solo grazie ad esso riusciamo a comunicare, costruire relazioni e creare comunità che ci hanno permesso di fare un considerevole salto evolutivo.

La stessa azione del comunicare è considerevole e sottovalutata: ci permette (non ironicamente!) di **leggere nelle menti** altrui.

Il vocabolario

Il vocabolario è il cuore pulsante della nostra lingua, **l'elemento comune** a qualsiasi dialetto.

Gli umani hanno un'ottima capacità di comunicare grazie a **set di lessemi** ed espressioni linguistiche che solitamente sono organizzate tramite **strutture gerarchiche**:

- Super-ordinate: categorie **inclusive**, generali, ampie
- Subordinate: categorie *specifiche* fatte di relazioni **iponimiche** con le precedenti

Basic/Advanced Level

Nel contesto sovraccitato si colloca la **nozione psico-linguistica** del *basic level*, ciò che, secondo la letteratura, possiede le seguenti caratteristiche:

- Livello *ottimale* di **economia cognitiva**
- Alto livello di **class-inclusion**
- Mediamente **generale**, culturalmente comune e *saliente*

Contributi della tesi

Identificazione di termini *basic/advanced + concrete/abstract* tramite due approcci computazionali:

- **Approccio testuale**: un large language model pre-addestrato e utilizzato generativamente
- **Approccio multi-modale**: una pipeline *text+image* che sfrutta reti neurali multi-modali stato dell'arte.

Inoltre, un ulteriore contributo è la creazione di un dataset di 500 parole *basic* e *advanced* esemplari grazie ad un *panel* di **dieci annotatori** language-learners e la **definizione** della nozione di **basicness**, basandosi sulla letteratura esistente sulla *concreteness*

How shall a thing be called? (1958)

Psychological Review
Vol. 65, No. 1, 1958

HOW SHALL A THING BE CALLED?

ROGER BROWN

Massachusetts Institute of Technology

The most deliberate part of first-language teaching is the business of telling a child what each thing is called. We ordinarily speak of *the name* of a thing as if there were just one, but in fact, of course, every referent has many names. The dime in my pocket is not only a *dime*. It is also *money*, a *metal object*, a *thing*, and, moving to subordinates, it is a *1952 dime*, in fact a *particular 1952 dime* with a unique pattern of scratches, discolorations, and smooth places. When such an object is named for a very young child how is it called? It may be named *money* or *dime* but probably not *metal object*, *thing*, *1952 dime*, or *particular 1952 dime*. The dog out on the lawn is not only a *dog* but is also a *boxer*, a *quadruped*, an *animate being*; it is the *landlord's dog*, named *Prince*. How will it be identified for a child? Sometimes it will be called a *dog*, sometimes *Prince*,

ing. It predicts the choice of *dime* over *metal object* and *particular 1952 dime*.

Zipf (10) has shown that the length of a word (in phonemes or syllables) is inversely related to its frequency in the printed language. Consequently the shorter names for any thing will usually also be the most frequently used names for that thing, and so it would seem that the choice of a name is usually predictable from either frequency or brevity. The monosyllables *dog* and *Prince* have much higher frequencies according to the Thorndike-Lorge list (8) than do the polysyllables *boxer*, *quadruped*, and *animate being*.

It sometimes happens, however, that the frequency-brevity principle makes the wrong prediction. The thing called a *pineapple* is also *fruit*. *Fruit* is the shorter and more frequent term, but adults will name the thing *pineapple*.

Roger Brown pone una *domanda fondamentale*:

Come facciamo a scegliere un **termine appropriato** per un **conetto** scelto?

Il basic di Brown

Individua **4 caratteristiche fondamentali** che devono condividere le parole basic:

- Brevi
- Concrete
- Facili da pronunciare
- Frequentemente utilizzate tra le disponibili per un certo concetto

Il nostro basic

Il basic di Brown

Individua **4 caratteristiche fondamentali** che devono condividere le parole basic:

- Brevi
- Concrete
- Facili da pronunciare
- Frequentemente utilizzate tra le disponibili per un certo concetto

Il nostro basic

- Strumenti per la **sopravvivenza sociale**

Il basic di Brown

Individua **4 caratteristiche fondamentali** che devono condividere le parole basic:

- Brevi
- Concrete
- Facili da pronunciare
- Frequentemente utilizzate tra le disponibili per un certo concetto

Il nostro basic

- Strumenti per la **sopravvivenza sociale**
- Brevi, facili da pronunciare *verbalmente* e da concettualizzare

Il basic di Brown

Individua **4 caratteristiche fondamentali** che devono condividere le parole basic:

- Brevi
- Concrete
- Facili da pronunciare
- Frequentemente utilizzate tra le disponibili per un certo concetto

Il nostro basic

- Strumenti per la **sopravvivenza sociale**
- Brevi, facili da pronunciare *verbalmente* e da concettualizzare
- Le **prime parole** che vengono alla mente quando si parla di un certo *topic*

Il basic di Brown

Individua **4 caratteristiche fondamentali** che devono condividere le parole basic:

- Brevi
- Concrete
- Facili da pronunciare
- Frequentemente utilizzate tra le disponibili per un certo concetto

Il nostro basic

- Strumenti per la **sopravvivenza sociale**
- Brevi, facili da pronunciare *verbalmente* e da concettualizzare
- Le **prime parole** che vengono alla mente quando si parla di un certo *topic*
- Facilmente traducibili in una **chiara immagine mentale**

Basic

- House, Book, Chair, Table
- Orange, Apple
- Dog, Cat
- Thing, Fact
- Justice, Fun

Advanced

- Mansion, Novel, Recliner, Board
- Tangerine, Granny Smith
- Chihuahua, Maine coon
- Artifact, Evidence
- Reprisal, Amusement

Utilizzo del basic level

- Utili per chi **impara una nuova lingua**: comprensione, produzione e conversazione
- Utilizzo nella **UI** di un prodotto
- Utilizzo in **automatic recommenders**, generatori di **sommari** ed **information extraction**
- Utilizzo per **text simplification** al fine di trattare *DSA* come la *Dislessia*

Attention Is All You Need

Aleksander
Krikunov
Google Brain
akrikunov@springer.com

Susan Wonnacott
Google Brain
swon@springer.com

Niki Parmar
Google Research
nikip@springer.com

Julia Chukharevich
Google Research
jchukh@springer.com

Elaine Jagger
Google Research
elaine.jagger@springer.com

Matthew W. Johnson¹
University of Toronto
mjohnson@cs.toronto.edu

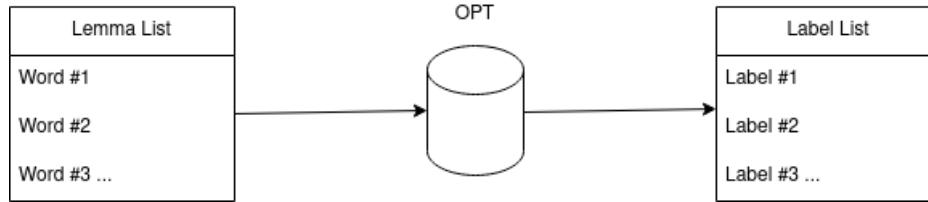
Kathleen Kucher
Google Brain
kucher@springer.com

Task sul livello testuale

OPT

Text-Based Pipeline

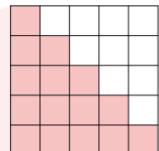
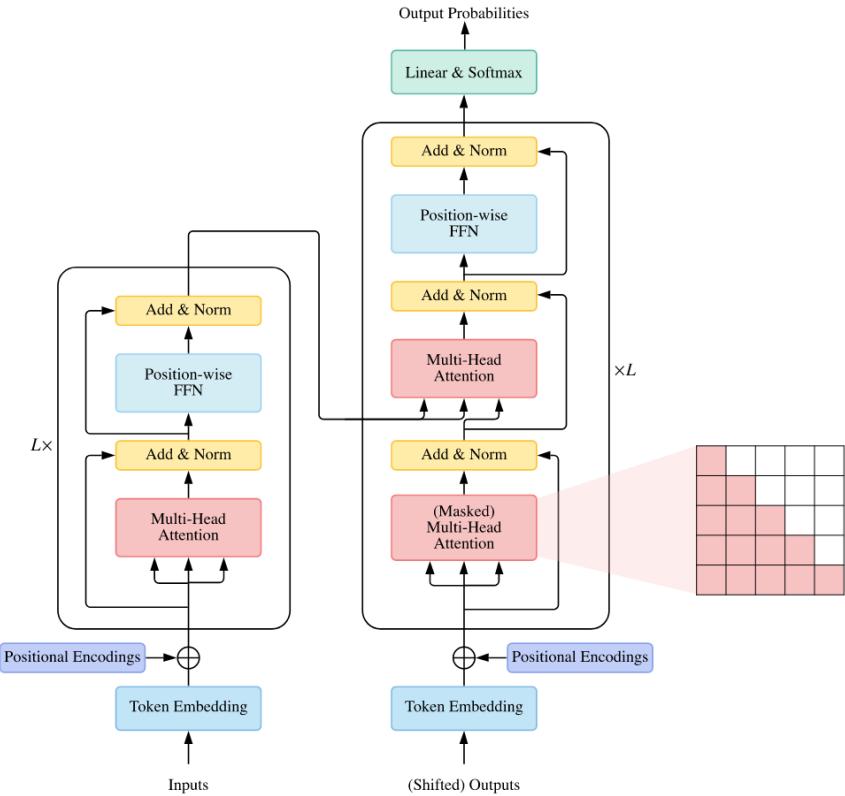
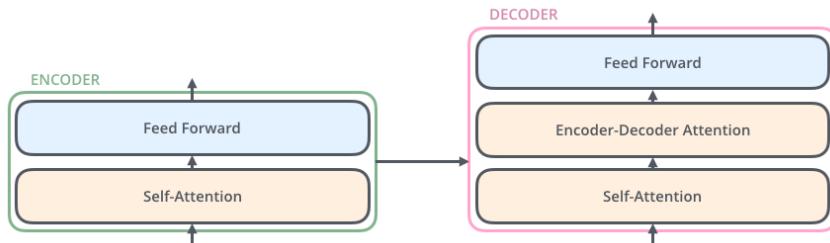
La prima pipeline utilizza *OPT*, un large language model generativo **transformer-based**.



Transformers

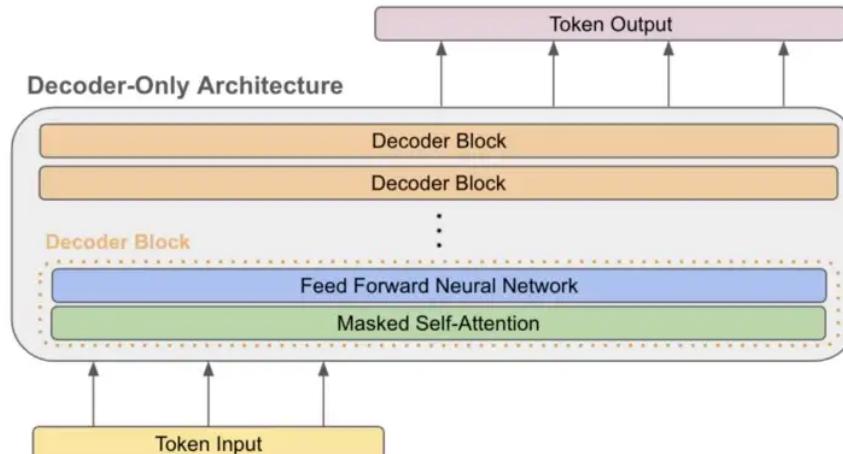
Rete neurale basata su uno **stack encoder-decoder**, adatta per dati di tipo testuale.

- **Encoder:** processa la sequenza di input (una serie di token) e restituisce **embeddings** (una rappresentazione contestuale verso il decoder) utilizzando layer di **self-attention**, **normalization** e **FFN**.
- **Decoder:** prende gli embeddings in input e ha come output finale una **distribuzione probabilistica** sul singolo token.



Architettura OPT

Suite di decoder-only pre-trained transformers: architettura **auto-regressiva** semplice, composta solo da decoder impilati. Utilizzato unicamente per scopi **generativi**.



I LLMs sono Next Token Predictors

- **Obiettivo:** prevedere il **token successivo** in una sequenza di parole, migliorando la coerenza e la comprensione del testo generato
- **Funzionamento:**
 - Viene utilizzata la *multi-headed attention* per catturare informazioni da contesti diversi
 - Calcola la **probabilità** di ciascun token candidato nel **vocabolario**
 - **Seleziona il token con la probabilità più alta** come previsione successiva

I LLMs sono Next Token Predictors

- **Obiettivo:** prevedere il **token successivo** in una sequenza di parole, migliorando la coerenza e la comprensione del testo generato
- **Funzionamento:**
 - Viene utilizzata la *multi-headed attention* per catturare informazioni da contesti diversi
 - Calcola la **probabilità** di ciascun token candidato nel **vocabolario**
 - **Seleziona il token con la probabilità più alta** come previsione successiva

Federico

I LLMs sono Next Token Predictors

- **Obiettivo:** prevedere il **token successivo** in una sequenza di parole, migliorando la coerenza e la comprensione del testo generato
- **Funzionamento:**
 - Viene utilizzata la *multi-headed attention* per catturare informazioni da contesti diversi
 - Calcola la **probabilità** di ciascun token candidato nel **vocabolario**
 - **Seleziona il token con la probabilità più alta** come previsione successiva

Federico prenderà

I LLMs sono Next Token Predictors

- **Obiettivo:** prevedere il **token successivo** in una sequenza di parole, migliorando la coerenza e la comprensione del testo generato
- **Funzionamento:**
 - Viene utilizzata la *multi-headed attention* per catturare informazioni da contesti diversi
 - Calcola la **probabilità** di ciascun token candidato nel **vocabolario**
 - **Seleziona il token con la probabilità più alta** come previsione successiva

Federico prenderà come

I LLMs sono Next Token Predictors

- **Obiettivo:** prevedere il **token successivo** in una sequenza di parole, migliorando la coerenza e la comprensione del testo generato
- **Funzionamento:**
 - Viene utilizzata la *multi-headed attention* per catturare informazioni da contesti diversi
 - Calcola la **probabilità** di ciascun token candidato nel **vocabolario**
 - **Seleziona il token con la probabilità più alta** come previsione successiva

Federico prenderà come voto

I LLMs sono Next Token Predictors

- **Obiettivo:** prevedere il **token successivo** in una sequenza di parole, migliorando la coerenza e la comprensione del testo generato
- **Funzionamento:**
 - Viene utilizzata la *multi-headed attention* per catturare informazioni da contesti diversi
 - Calcola la **probabilità** di ciascun token candidato nel **vocabolario**
 - **Seleziona il token con la probabilità più alta** come previsione successiva

Federico prenderà come voto di

I LLMs sono Next Token Predictors

- **Obiettivo:** prevedere il **token successivo** in una sequenza di parole, migliorando la coerenza e la comprensione del testo generato
- **Funzionamento:**
 - Viene utilizzata la *multi-headed attention* per catturare informazioni da contesti diversi
 - Calcola la **probabilità** di ciascun token candidato nel **vocabolario**
 - **Seleziona il token con la probabilità più alta** come previsione successiva

Federico prenderà come voto di laurea

I LLMs sono Next Token Predictors

- **Obiettivo:** prevedere il **token successivo** in una sequenza di parole, migliorando la coerenza e la comprensione del testo generato
- **Funzionamento:**
 - Viene utilizzata la *multi-headed attention* per catturare informazioni da contesti diversi
 - Calcola la **probabilità** di ciascun token candidato nel **vocabolario**
 - **Seleziona il token con la probabilità più alta** come previsione successiva

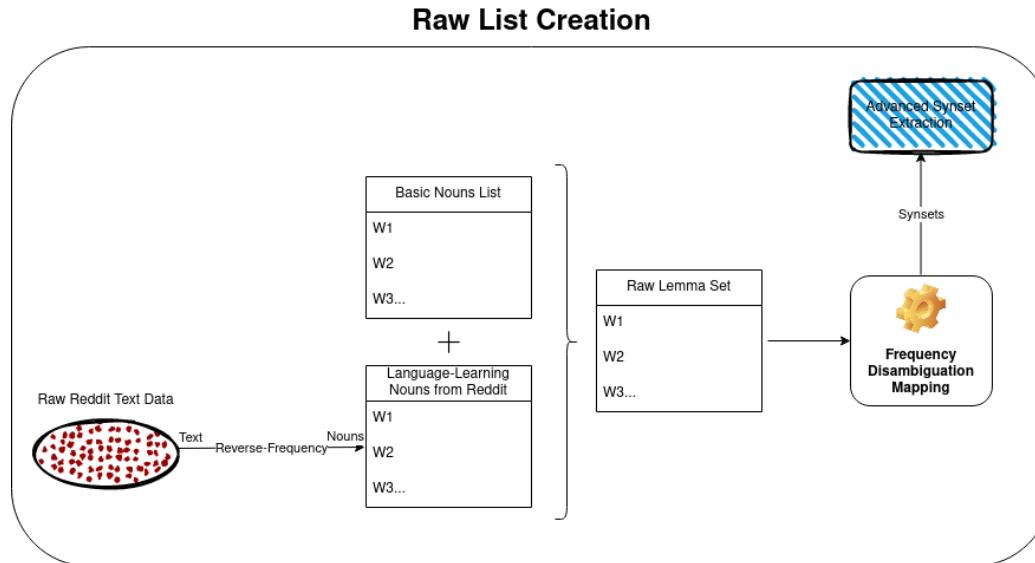
Federico prenderà come voto di laurea ...

Come classificare con un modello generativo?

- **Creazione** di una *basic raw list*
- **Filtraggio** tramite *OPT*
- **Estrazione** di termini Advanced
- **Raffinamento** del dataset finale

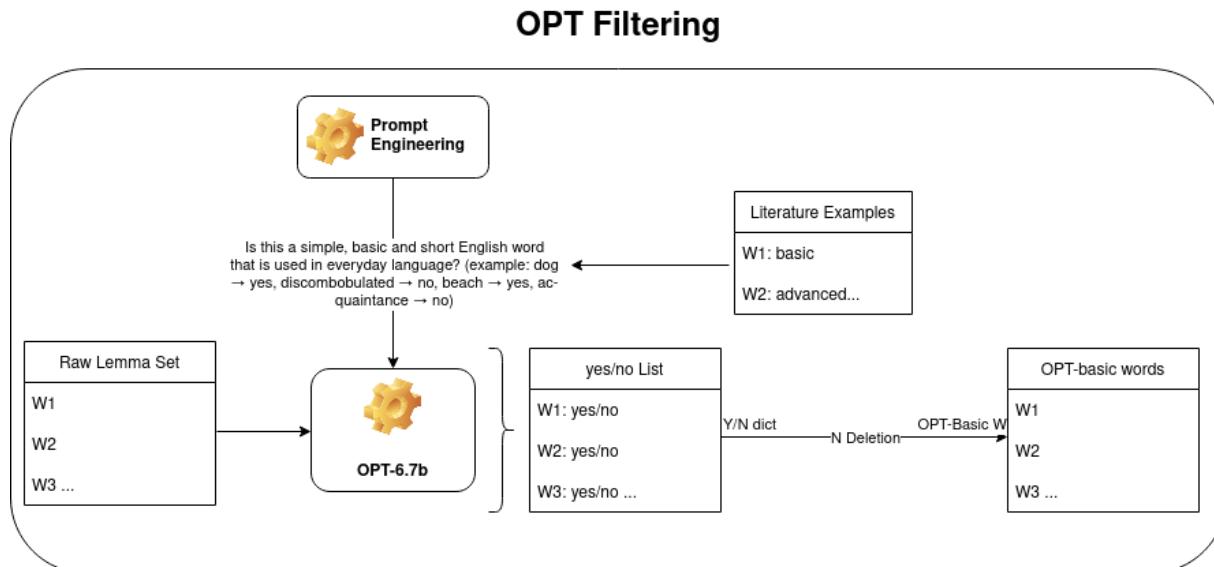
Creazione della basic raw list

- **Input:** termini estratti da liste di termini *semplici* (e.g. Ogden) e da testi di language-learners su Reddit
- **Output:** set iniziale di termini da *filtrare* usando **OPT** e, contemporaneamente, i loro synset associati (*best-frequency*) per l'estrazione di **advanced**



Filtraggio tramite OPT

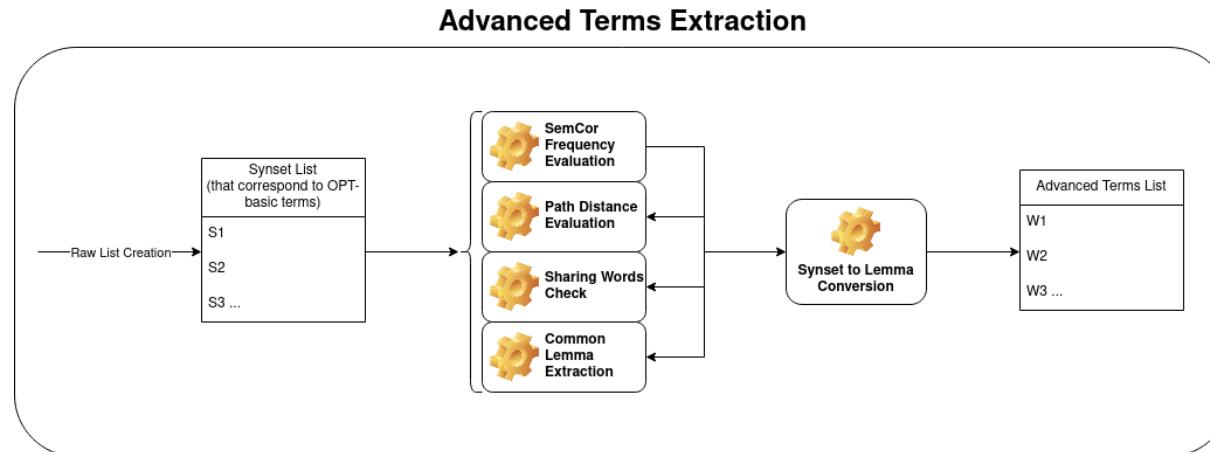
- **Input:** il set iniziale di termini
- **Output:** una lista di parole **OPT-basic**



Estrazione di termini advanced

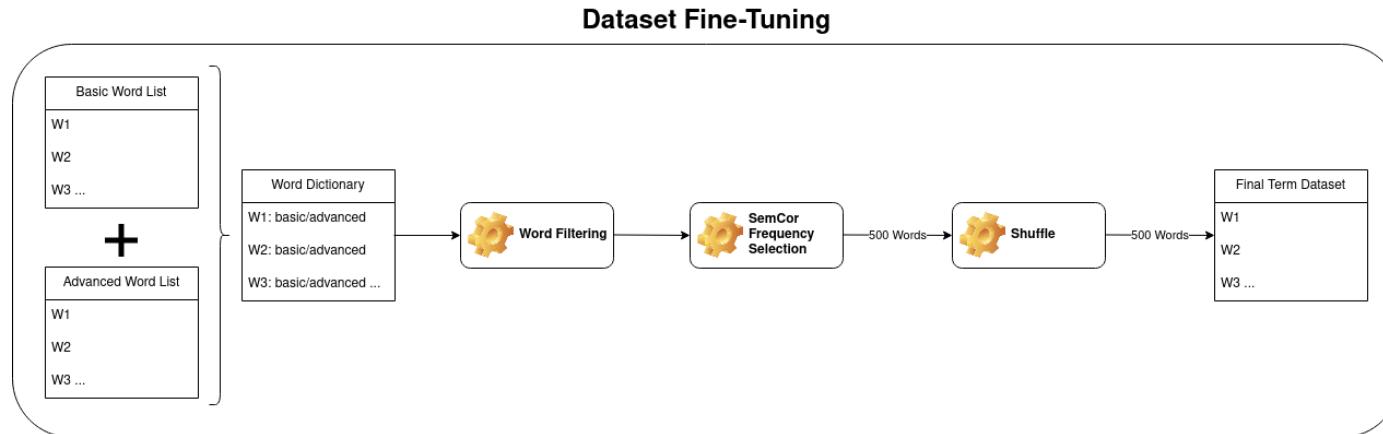
Partendo dai synset basic vengono fatti controlli:

- Frequenza significativa in **SemCor**, un corpus annotato
- **Path Distance** appropriata
- **Nessuna parola condivisa** tra il *synset* e l'*iponimo*
- Non devono esserci parole basic nella advanced list



Dataset fine-tuning

Vengono prodotti **500 termini**, 250 sono *OPT-basic* e 250 sono *OPT-advanced*. Le due liste (basic+advanced) vengono **combinate**, **filtrate** per rimuovere *parole offensive* o dannose, selezionate in base alla loro **frequenza** e poi **mischiare** per produrre il dataset finale da somministrare agli utenti.



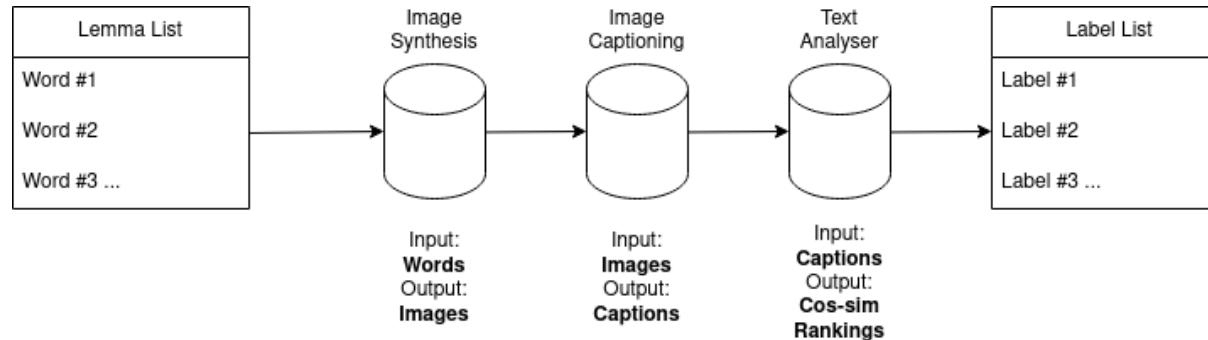
Task sul livello visivo

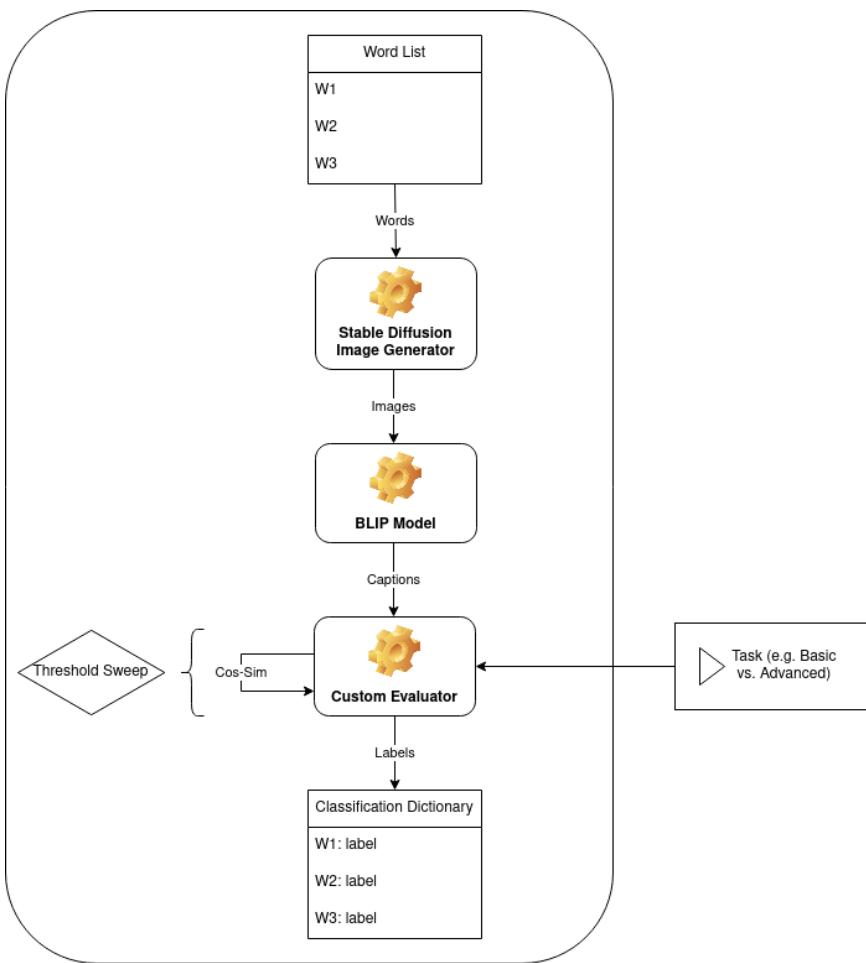
stableKnowledge



Pipeline multi-modale

- *text-to-image*: Stable Diffusion
- *image-to-text*: BLIP
- *semantic text evaluation*: SBERT



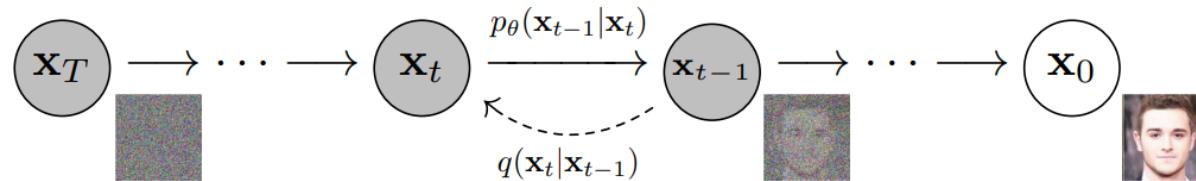


Come mai abbiamo bisogno del livello visivo?

Ipotesi: un sistema che possiede l'abilità di riconoscere l'informazione dal livello testuale al livello visivo (e viceversa), ha una conoscenza profonda del suo contenuto, in maniera similare a come l'uomo conosce il mondo.

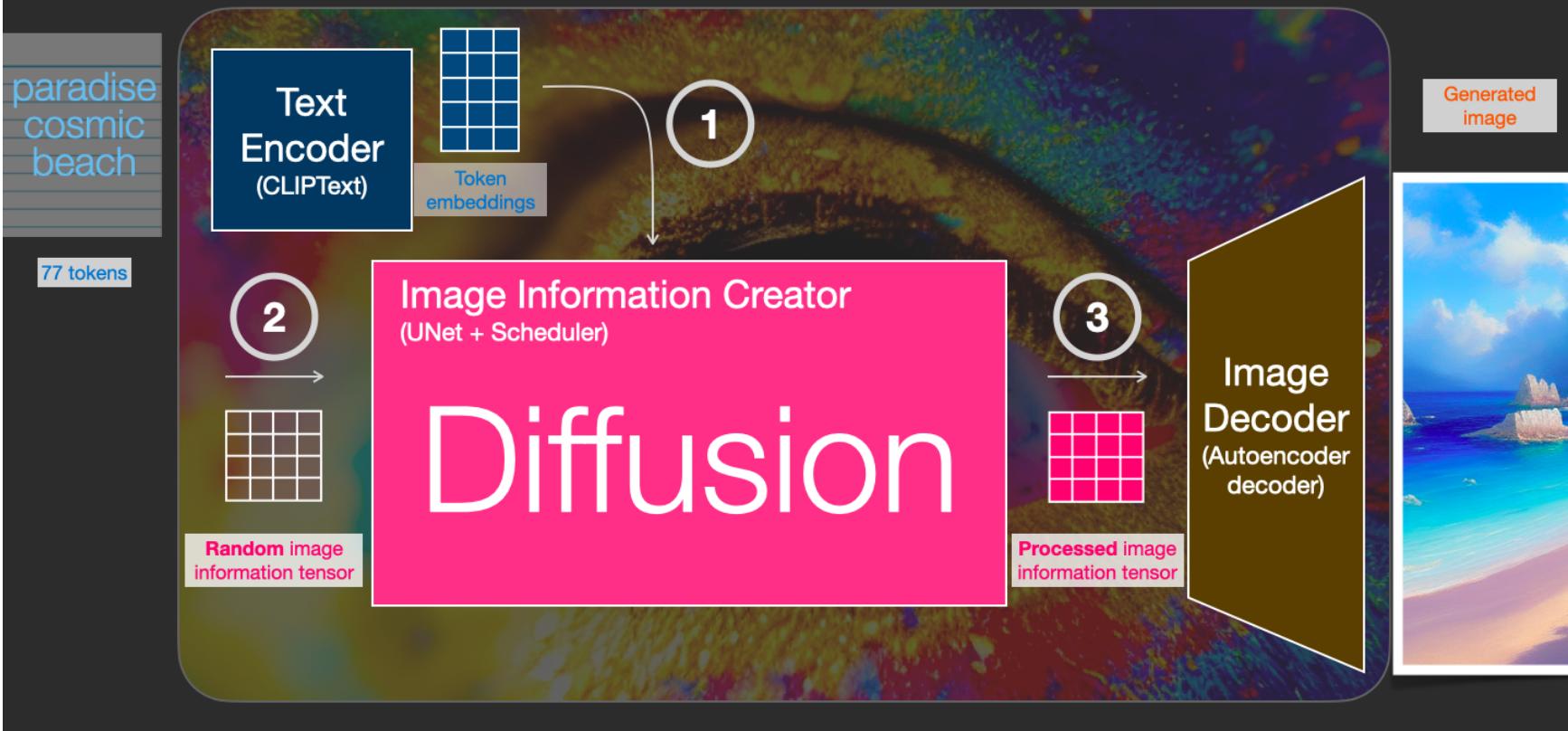
Generare immagini

- Viene utilizzata la **diffusione**: un **processo a doppia passata** per allenare modelli a partire da enormi dataset di immagini
 - **Forward pass**: viene *distrutta* la struttura dei dati aggiungendo del **rumore Gaussiano** all'immagine attraverso una *catena di Markov*. I dati vengono *assorbiti* nello spazio latente.
 - **Backward pass**: si cerca di re-imparare i dati di origine andando a progressivamente **rimuovere il rumore dalle immagini**, navigando nello spazio latente e generando nuovi dati





Stable Diffusion



Interrogare immagini

Viene utilizzato **BLIP**, un sistema multi-modale capace di **generare delle descrizioni** delle immagini presentate. Gli autori del paper hanno scelto di utilizzare un **ViT** (Vision Transformer) e un **MED** (Multi-modal mixture of Encoder-Decoder).

- **ViT**: scomponete l'immagine di partenza in **features** e trasforma le features in una sequenza di **embeddings**
- **MED**: produce degli stati dati in pasto ad un Transformer, che fa **image captioning**. MED opera come **Unimodal Encoder** (ITC Loss) e **Image-grounded text enc/decoder** (ITM+LM Loss)

Secondo gli autori BLIP è "*an unholy concoction of many different things, in one, trained jointly*"

Valutare immagini

Valutiamo se la descrizione si avvicina al lemma originale utilizzando **SBERT**, una rete *siamese* (due reti identiche in training, comparate in testing) BERT-based che utilizza un layer di *pooling* per generare degli **embeddings** utilizzati per confrontare testi in uno spazio semantico utilizzando la **cosine-similarity**.

Abbiamo costruito un componente *custom* per adattare il task di classificazione selezionato

Classificazione

Una volta raccolte le *cosine-similarities* di tutte le immagini, si fa un **threshold sweep** per verificare quale sia il limite superiore di classificazione.

e.g.: se x è il threshold, allora: $\forall y < x, y$ basic, advanced altrimenti.

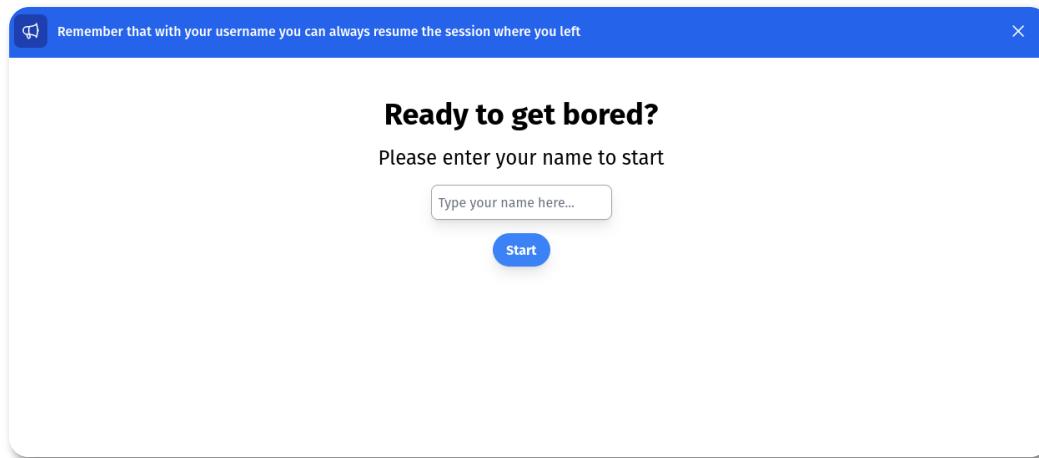
Basic and concrete input: "bird"	Basic and abstract input: "process"
Generated images (sample of two out of five) 	Generated images (sample of two out of five) 
Generated caption (left): <i>a brown bird with black wings and red feet sitting on top of a tree</i> Generated caption (right): <i>a red head bird with black and grey stripes</i>	Generated caption (left): <i>a painting of a man with his face draw</i> Generated caption (right): <i>the woman is looking at herself through the door</i>
Mean Cosine Similarity: 0.6	Mean Cosine Similarity: 0.1

Il task di annotazione

Per comparare i dati classificati dai metodi automatici con uno **standard** abbiamo creato un tool personalizzato e sottoposto **10 second-language-learners** al task di annotazione di **basic vs. advanced** sulle *500 parole*.

Tra i dati interessanti raccolti menzioniamo:

- Statistiche sul **tempo**
- Parole classificate come **difficili da annotare**
- **Agreement** tra gli annotatori



Basic or Advanced?

Word: war

BASIC ADVANCED

Was this hard to evaluate?

Agreement tra gli annotatori

Abbiamo ottenuto un **alto** inter-annotation agreement, misurato con il **Coefficiente k di Cohen**, la cui formula è

$$k = \frac{p_o - p_e}{1 - p_e}$$

Dove:

- p_o è l'agreement osservato relativo tra gli annotatori
- p_e è la probabilità ipotetica di agreement *casuale*

Si misura in una scala che va da **0** (nessun agreement) a **1** (agreement perfetto).

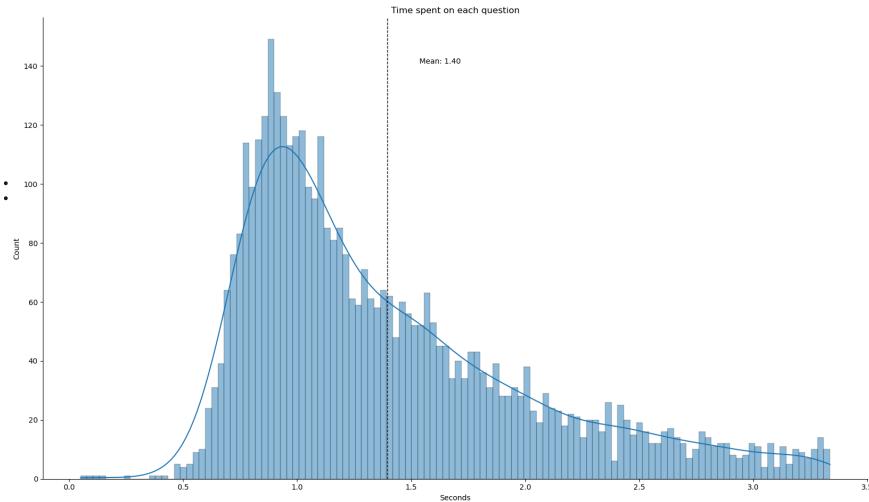
Cohen's Kappa	Interpretation
0	No agreement
0.10 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Near perfect agreement
1	Perfect agreement

0.71

agreement tra annotatori, che indica un task con **ottime guidelines**

Fatti interessanti

- Gli annotatori hanno impiegato **più tempo**
(80% in più) a valutare parole basic piuttosto
che quelle advanced
- Il tempo medio è stato di **1.40 secondi**
- Le parole più difficili da classificare sono state:
 - Parole composte da più basic: **vitamin pill**
 - Parole prese da altre lingue: **avenue**
 - Parole corte ma con un significato
complesso: **kin**



Basicness

Dalle analisi è risultato che la classificazione **basic vs. advanced** non è una **proprietà binaria**, ma deve essere posizionata su una scala misurabile.

Da questa ipotesi è nata la **basicness**, una misura similare alla *concreteness* che **segnala quanto è basic un certo termine**, basandosi sullo **split** nell'annotazione del lemma stesso.

Annotators split	#basic	#adv	#total
<i>low agreement</i>			
oooooo ooooo	35	35	132
oooooooo oooo	35	27	
<i>medium agreement</i>			
ooooooooo ooo	48	43	203
oooooooooo oo	52	60	
<i>high agreement</i>			
oooooooooooo o	35	48	165
oooooooooooooo	41	41	

Risultati sul task basic vs. advanced

	<i>OPT model (k = 0.63)</i>			<i>stableKnowledge (k = 0.21)</i>		
	Precision	Recall	F1	Precision	Recall	F1
<i>basic</i>	0.82	0.81	0.82	0.61	0.59	0.60
<i>advanced</i>	0.81	0.82	0.82	0.59	0.61	0.60

Una scoperta interessante

Investigando sui dati raccolti abbiamo scoperto che **termini basic e concreti sono meglio riconosciuti dal livello testuale.**

Questo ci ha portato a ipotizzare che l'**image level** fosse più adatto a classificare la **concreteness**.

Una scoperta interessante

Investigando sui dati raccolti abbiamo scoperto che **termini basic e concreti sono meglio riconosciuti dal livello testuale.**

Questo ci ha portato a ipotizzare che l'**image level** fosse più adatto a classificare la **concreteness**.

L'ipotesi è stata confermata dal nostro *secondo esperimento*: **concrete vs. abstract**

- Presa una lista di concetti dal database di **MRC**, con score **[0, 700]**
- Classificati tutti i sostantivi nella lista con i nostri due metodi (con fine-tuning sulla concreteness)
- Analizzati i risultati

Risultati sul task concrete vs. abstract

	<i>OPT model (k = 0.27)</i>			<i>stableKnowledge (k = 0.57)</i>		
	Precision	Recall	F1	Precision	Recall	F1
<i>abstract</i>	0.63	0.40	0.49	0.77	0.85	0.81
<i>concrete</i>	0.70	0.85	0.77	0.82	0.72	0.76

Questi risultati indicano che quando si tratta di differenziare concetti concreti da astratti, i due metodi performano **al contrario** rispetto al task precedente!

Risultati sul task concrete vs. abstract

	<i>OPT model (k = 0.27)</i>			<i>stableKnowledge (k = 0.57)</i>		
	Precision	Recall	F1	Precision	Recall	F1
<i>abstract</i>	0.63	0.40	0.49	0.77	0.85	0.81
<i>concrete</i>	0.70	0.85	0.77	0.82	0.72	0.76

Questi risultati indicano che quando si tratta di differenziare concetti concreti da astratti, i due metodi performano **al contrario** rispetto al task precedente!

L'origine di questa discrepanza è nell'architettura e nel **training set** dei modelli utilizzati:

- OPT è **text-based**, ovvero, addestrato in maniera supervisionata su enormi quantità di dati testuali
- SD è **image-based**, ovvero, addestrato su **immagini e captions**

Sum-up

- Proposta una nuova nozione di "**basicness**" ispirata alla letteratura sulla concreteness
- Creato un'ampia **wordlist inter categoria** di parole basic/advanced
- Sviluppata una metodologia **human-in-the-loop** con 10 annotatori con *alto agreement*
- Creati metodi per la classificazione di elementi terminologici utilizzando reti neurali stato dell'arte per **CV** e **NLP**
- Mostrata una pipeline image-based che combina **text-to-image** e **image-to-text** per scopi NLP, novità assoluta in letteratura

Conclusion: Future Work

- **Scoperte secondarie:** Gli approcci basati su immagini funzionano meglio per i termini concreti e astratti ma non per la classificazione basic/advanced. È stata formulata un'ipotesi ma sono necessarie ulteriori ricerche.
- Le pipeline multimodali potrebbero migliorare gli attuali Large Language Model che replicano solo la **rete linguistica** e mancano di vera intelligenza. Le affermazioni secondo cui i LLMs hanno proprietà umane come la Teoria della Mente richiedono un esame critico.
- Alcune immagini generate hanno mostrato una **qualità "uncanny"** che richiede ulteriori approfondimenti per risolvere il problema mantenendo un'alta qualità.

Sidenote Finale

- I ricercatori dovrebbero avere una **prospettiva equilibrata** sui LLMs e sugli LDMs. Non dovrebbero essere temuti, glorificati o eccessivamente pubblicizzati. Sono strumenti per aiutare gli umani ma non sono intelligenti o in grado di pensare o immaginare come gli umani. Un'enfasi eccessiva potrebbe portare a un nuovo **"AI Winter"**
- Sebbene le aziende promuovano i progressi nell'AI, le università dovrebbero competere con loro per **promuovere modelli aperti** e ricerche per una vera apertura. La ricerca condotta solo dalle aziende potrebbe mancare della **necessaria trasparenza e collaborazione** per condividere i risultati con la comunità di ricerca in generale.



Grazie per l'attenzione

