Relazione Esercizio 2

Similarità Lessicale/Semantica

SimLex

Per il calcolo della similarità lessicale (simlex) è stato adottato un approccio statistico che misura la **sovrapposizione dei termini** tra diverse definizioni dello stesso concetto. A tal scopo, per prima cosa ciascuna definizione viene sottoposta ad alcune operazione di pre-processing del testo, tra cui:

- tokenization, per ottenere i singoli termini della definizione.
- stopword and punctuation filtering, per eliminare il rumore.
- stemming, per otterene una forma normalizzata dei termini.

Successivamente, le definizioni così processate vengono iterate per calcolare il valore di similarità lessicale (simlex) in ciascuna categoria. In particolare, il calcolo avviene considerando l'insieme di termini di ciascuna definizione, prese a due a due, e calcolando la percentuale di termini condivisi. Questo valore è ottenuto dividendo il numero di termini comuni tra le due definizioni per il numero totale di termini unici presenti in entrambe.

$$\text{simlex} = \frac{\texttt{tokens(def1)} \cap \texttt{tokens(def2)}}{\texttt{tokens(def1)} \cup \texttt{tokens(def2)}}$$

– Risultati

Questi singoli valori vengono poi aggregati calcolando la loro media all’interno di ciascuna categoria, ottenendo i seguenti risultati:

Pantalone[CG]	Microscopio[CS]	Pericolo[AG]	Euristica[AS]
22.91	18.43	15.36	9.14

Dall'analisi dei risultati si evidenzia come la similarità lessicale misurata presenti valori relativamente bassi per tutte le categorie. Questo fenomeno può essere attribuito a diversi fattori. In primo luogo, la semplicità della tecnica di calcolo utilizzata, basata esclusivamente sulla sovrapposizione di termini, non riesce a catturare le relazioni semantiche più profonde tra le parole. In secondo luogo, la grande varietà di linguaggio impiegata nelle definizioni, combinata con un numero limitato di esempi, impatta significativamente sui risultati, poiché anche piccole variazioni lessicali vengono interpretate come differenze sostanziali, soprattutto su definizioni molto brevi.

È interessante notare come i valori di similarità seguano un pattern decrescente che riflette la natura dei concetti: i concetti concreti e generali (Pantalone [CG]) mostrano la similarità più alta (22.91%), mentre i valori diminuiscono progressivamente quando si passa a concetti concreti ma specifici (Microscopio

[CS], 18.43%), per arrivare ai minimi per i concetti astratti, sia generali (Pericolo [AG], 15.36%) che specifici (Euristica [AS], 9.14%).

Questo trend conferma l'ipotesi che i concetti concreti siano più facili da definire con un vocabolario più condiviso e standardizzato, mentre i concetti astratti, e specialmente quelli specifici, vengono espressi con terminologia più varia e personalizzata, conseguenza della difficoltà pratica nel definire tali concetti.

Infine, mostriamo quali sono le definizioni che hanno ottenuto il punteggio maggiore per ciascuna categoria:

Categoria	Definizione 1	Definizione 2	Punteggio SimLex
Pantalone[CG]	[P20] Capo di abbigliamento per la parte inferiore del corpo	P[36] capo di abbigliamento indossabile nella parte inferiore del corpo	0.86
Microscopio[CS]	[P20] Strumento per osservare oggetti di piccole dimensioni non visibili ad occhio nudo	[P36] strumento scientifico per osservare oggetti non visibili ad occhio nudo	0.66
Pericolo[AG]	[P8] situazione che compromette la sicurezza di un soggetto	[P9] situazione che minaccia la sicurezza di un soggetto	0.6
Euristica[AS]:	[P5] Funzione che stima la distanza dallo stato attuale allo stato obiettivo.	[P31] Funzione utilizzata per stimare la distanza dallo stato attuale al goal.	0.625

SimSem

Per quanto riguarda la similarità semantica delle definizioni, è stato utilizzato un approccio distribuzionale, trasformando le definizioni in embeddings all'interno di uno spazio vettoriale. A questo scopo è stato utilizzato il modello di embeddings General Text Embeddings (GTE), nella sua versione small, che permette di trasformare una sequenza di token (una frase) in una rappresentazione vettoriale.

Ottenuto l’insieme di embeddings, la similarità è stata quindi calcolata misurando la distanza tra i vettori corrispondenti a ciascuna definizione. In particolare è stata utilizzata la Cosine Similarity che considera l'angolo tra i vettori delle definizioni.

— Risultati e Confronto

Pantalone[CG]	Microscopio[CS]	Pericolo[AG]	Euristica[AS]
87.83	88.24	86.35	85.15

I valori di similarità ottenuti tramite questo approccio risultano essere molto più elevati. Questo è dovuto al fatto che una rappresentazione distribuzionale riesce a catturare la semantica espressa da ciascuna definizione, indipendentemente dai termini specifici utilizzati.

È interessante notare come il divario tra le diverse categorie si sia notevolmente ridotto rispetto ai risultati della similarità lessicale. Questo suggerisce che,

sebbene il vocabolario utilizzato possa variare significativamente (come evidenziato dai bassi valori di SimLex), il significato sottostante rimane relativamente coerente all'interno di ciascuna categoria di concetti. Questa osservazione porta alla conclusione che la difficoltà nel definire concetti astratti e specifici si manifesti principalmente a livello lessicale, mentre a livello semantico profondo, le definizioni tendono a convergere verso significati simili.

Infine, seppur in maniera molto più lieve, si conferma l'andamento decrescente dei valori tra le categorie. Questo rafforza l'evidenza della difficoltà nel definire in modo condiviso termini più astratti e specifici.