Relazione Esercizio 1

Integrare risorse eterogenee: WordNet + Embeddings

Motivazioni

Uno dei limiti dei dizionari elettronici, come WordNet, è la loro natura statica e discreta: relazioni semantiche come iperonimia, iperonimia, sinonimia, ecc. non sempre riescono a catturare informazioni di vicinanza semantica tra termini. Questo genere di informazioni può essere estratta però da modelli modelli distribuzionali basati su word e sentence embeddings, i quali sono in grado di cogliere similarità semantiche (o di utilizzo) e associazioni implicite che spesso non sono esplicitamente codificate in risorse lessicali.

Progettazione

L'idea è quindi di costruire inizialmente uno spazio semantico in cui ogni synset di WordNet viene mappato in un vettore di embedding. Questo spazio sarà poi utilizzato per identificare gli elementi semanticamente più vicini a un dato synset di query, arricchendo così il contenuto informativo portato da quest'ultimo.

Implementazione

Per la creazione dell’insieme di embeddings stato creato un corpus formato dalle definizioni di ciascun synset all’interno di wordnet. Tutti questi testi sono stati forniti al modello General Text Embeddings (GTE). Per motivi di efficienza, i vettori risultati sono stati naturalmente serializzati all’interno di un file salvato localmente.

Come anticipato, questo spazio di rappresentazione vettoriale viene utilizzato per individuare i k elementi semanticamente più vicini a un determinato synset. Il processo di ricerca avviene attraverso il calcolo della cosine similarity tra il vettore del synset di "query" e tutti gli altri vettori presenti nello spazio. L'algoritmo identifica e restituisce i k elementi con il valore di similarità più elevato, fornendo così un'estensione semantica contestuale che arricchisce l'informazione originale del synset.

Risultati

Per illustrare i risultati dell'integrazione dei due sistemi, sono state create due tabelle distinte:

- La prima tabella presenta alcune informazioni chiave fornite direttamente da WordNet sulla parola target.
- La seconda tabella mostra i top k synset semanticamente più vicini a quello della parola target, includendo la descrizione e il relativo grado di similarità.

Vediamo un esempio considerando come target il termine “computer”:

--- WORDNET RESULT FOR computer.n.01 ---

GLOSS (DEFINITION)	a machine for performing calculations automatically
LEMMAS	['computer', 'computing_machine', 'computing_device', 'data_processor', 'electronic_computer', 'information_processing_system']
HYPERNYMS	[Synset('machine.n.01')]
HYPONYMS	[Synset('server.n.03'), Synset('turing_machine.n.01'), Synset('number_cruncher.n.02'), Synset('analog_computer.n.01'), Synset('home_computer.n.01')]...

--- EMBEDDINGS RESULT FOR computer.n.01 ---

SYNSET	DESCRIPTION	SEMSIM
calculator.n.02	a small machine that is used for mathematical calculations	0.95
adder.n.02	a machine that adds numbers	0.92
calculator.n.01	an expert at calculation (or at operating calculating machines)	0.92
subtracter.n.02	a machine that subtracts numbers	0.92
number_cruncher.n.02	a computer capable of performing a large number of mathematical operations per second	0.91

Come possiamo osservare, la ricerca nello spazio degli embeddings ci consente di ottenere un insieme di synset semanticamente collegati a quello iniziale, fornendo uno spettro di informazioni significativamente più ampio, superando la rigida tassonomia di relazioni lessicali offerta dal solo WordNet.