Relazione Esercizio 3

Content To Form

Pre-Processing e Progettazione

— Traduzione delle Definizioni

Questa esercitazione affronta il problema della **ricerca onomasiologica**, cioè l'individuazione di un concetto partendo dalla sua definizione, il che si traduce praticamente nel processare tutte le definizioni raccolte al fine di individuare il synset (più probabile) a cui fanno riferimento e verificare se tale synset corrisponde al concetto “target” che la definizione dovrebbe esprimere. A tal scopo, è stato innanzitutto **necessario** implementare una funzione (**traduci_definizioni()**) che vada **tradurre tutte le definizioni** dall’italiano all’inglese, sfruttando le API di Google Translator.

— Estrazione del Genus

Per risolvere il task, l’idea è quella di sfruttare il modello di ***Genus-Differentia***, ottenendo quindi il “genus” di ciascuna definizione e ricercando tra gli iponimi di quest’ultimo il concetto iniziale espresso dalla definizione. L'ipotesi di base è che il genus corrisponda al termine più frequente all’interno di un insieme di definizioni. È stato quindi implementato il metodo **get_k_genus(definitions, k)** che raggruppa le definizioni di un concetto, le preprocessa per ridurre il rumore, e identifica i k termini più frequenti per determinare il genus. I genus ottenuti per le definizioni in questa esercitazione sono:

- Pantalone[CG]: indumento (garment)
- Microscopio[CS]: strumento (tool)
- Pericolo[AG]: situazione (situation)
- Euristica[AS]: problema (problem)

Questi risultati dimostrano che la tecnica implementata estrae in modo efficiente il genus utilizzato nelle definizioni, corrispondendo al termine più frequente nei primi tre casi. Come prevedibile, l'estrazione per il concetto di "Euristica[AS]" risulta più complessa, con *problema* come termine più frequente, mentre *"strategia"* appare solo al quinto posto.

Soluzione 1: Ricerca “Locale” ai Genus

[onomasiological_by_sentence_embeddings(definitions, k)]

[onomasiological_by_word_freq(definitions, k)]

Nel primo approccio alla soluzione di questo problema, la ricerca è stata focalizzata all’interno di un insieme di synset “candidati”, identificati esplorando tutti gli iponimi del genus e, per migliorare la robustezza, includendo anche gli iponimi dei sinonimi del genus presenti in WordNet.

Una volta costruito tale insieme, sono state utilizzate **analisi delle frequenze ed embeddings vettoriali** per identificare i k synset candidati più probabili. Il processo di ricerca onomasiologica viene considerato riuscito quando il synset target risulta presente tra questi candidati selezionati. Per i prossimi risultati è stato utilizzato un $k = 4$

Risultati ricerca by Word Frequency

- Pantalone[CG]: 12/39 guessed.
- Microscopio[CS]: 0/39 guessed.
- Pericolo[AG]: 0/39 guessed.
- Euristica[AS]: 0/39 guessed.

Risultati ricerca by Sentence Embeddings:

- Pantalone[CG]: 21/39 guessed.
- Microscopio[CS]: 0/39 guessed.
- Pericolo[AG]: 0/39 guessed.
- Euristica[AS]: 0/39 guessed.

Analizzando questi risultati, emerge che la "semplice" ricerca basata sulla frequenza dei termini non è affatto efficace. Ancora più rilevante è come una ricerca limitata agli iponimi del genus estratto dalle definizioni risulti fortemente insufficiente. Le cause di questo fenomeno risiedono nella struttura stessa delle definizioni, che non consentono di identificare il corretto "concetto generale" da cui derivano i concetti delle definizioni analizzate. Infatti:

- Per “Microscopio” è impossibile utilizzare gli iponimi del genus (strumento) perchè in WordNet “Microcope” è un iponimo di “magnifier” ovvero lente di ingrandimento. Termine che però non compare mai nelle definizioni considerate.
- Per “Pericolo” il Genus “Situazione” contiene effettivamente situazioni di pericolo ma in WordNet l’iperonimo giusto è “Condition”, “Status”. Anche in questo caso sarà quindi impossibile risalire al giusto synset.

Di seguito un esempio di output ottenuto per quanto riguarda la categoria di definizioni Concreto Generico - Pantalone nella ricerca Word Freq.

Definition ID	Possible Synsets	Result
P2	[Synset('trouser.n.01'), Synset('legging.n.01'), Synset('separate.n.02'), Synset('overgarment.n.01')]	**True**
P3	[Synset('legging.n.01'), Synset('overgarment.n.01'), Synset('shirt.n.01'), Synset('weeds.n.01')]	**False**
P4	[Synset('legging.n.01'), Synset('trouser.n.01'), Synset('overgarment.n.01'), Synset('camlet.n.01')]	**True**
…	…	…

Soluzione 2: Ricerca “Global”

`[onomasiological_by_sentence_embeddings(definitions, k)]`

È stato implementata anche una seconda tipologia di ricerca onomasiologica che in maniera globale va a ricercare il possibile concetto da associare alla definizione ricercando tra tutti quelli presenti all’interno di wordnet. In questo caso la ricerca si è basata esclusivamente sul confronto della similarità tra gli embeddings da ciascun synset, sfruttando la loro descrizione fornita da wordnet.
Risultati ($k = 4$):

- Pantalone[CG]: 0/39 guessed.
- Microscopio[CS]: 5/39 guessed.

- Pericolo[AG]: 0/39 guessed.
- Euristica[AS]: 1/39 guessed.

In questo caso le prestazioni peggiorano notevolmente, probabilmente a causa della grande dispersione di informazioni che si crea quando si mappano tutti i synset in uno spazio vettoriale. Tuttavia, concetti prima irraggiungibili come Microscopio ed Euristica, seppur in minima parte, vengono comunque individuati.

Definition ID	**Possible Synsets**	**Result**
P21	[Synset('anatomize.v.02'), Synset('fine-tooth_comb.n.02'), Synset('microscope.n.01'), Synset('microcosm.n.01')]	**True**
P1	[Synset('scientific_instrument.n.01'), Synset('optical_bench.n.01'), Synset('microtome.n.01'), Synset('telemeter.n.01')]	**False**
P29	[Synset('microscope.n.01'), Synset('ultramicroscope.n.01'), Synset('telescopic.s.02'), Synset('farsighted.a.01')]	**True**
…	…	…