ANNO ACCADEMICO 2024/2025

Tecnologie del Linguaggio Naturale

Esercitazioni

Luca Barra

UNIVERSITÀ
DI TORINO

DIPARTIMENTO DI INFORMATICA

1 WordNet to ConceptNet Mapping

L'idea alla base di quest'esercitazione è: creare un collegamento tra WordNet (che è un database lessicale contente relazioni semantiche[1]) con ConceptNet (un database di *common knowledge*). Unendo entrambe le risorse se ne può creare una nuova in grado di massimizzare i punti di forza di entrambe, per esempio si può migliorare l'accuratezza di task come la word sense disambiguation (WSD) migliorando l'accuratezza di algoritmi come quello di Lesk.

Listing 1.1: The new resource obtain from the combination of WordNet and ConceptNet

```
return {
  "synsets": [
    {
      "name": syn.name(),
      "definition": syn.definition(),
    }
    for syn in synsets
  ],
  "relations": relations
}
```

[1] Sinonimia, iperonimia, iponimia, etc

2 Definitions

2.1 Definition Complexity

L'obiettivo di quest'esercitazione è quello di far ragionare su quanto sia difficile creare delle definizioni che siano univoche e condivise. Per far ciò sono state scelte 4 parole e per ciascuna di esse diverse persone hanno fornito delle definizioni:

- Concetti concreti:
  - Generico: *pantalone*.
  - Specifico: *microscopio*.
- Concetti astratti:
  - Generico: *pericolo*.
  - Specifico: *euristica*.

Passi preliminari:

- Le definizioni sono state tradotte in inglese: questo perché così facendo è possibile utilizzare WordNet (questo riguarda la prossima esercitazione *content to form*).
- Le definizioni sono state lemmatizzate, tokenizzate e sono state rimosse le stopwords.

2.1.1 SimLex

Per valutare la similarità lessicale si sono contati i termini in comune tra le definizioni prese a due a due e si è fatta la media.

Table 2.1: Similarità Lessicale.

Termine	Similarità Lessicale
Pantalone	0.1950
Microscopio	0.1519
Pericolo	0.1185
Euristica	0.2276

Considerazioni:

- Concreto/Astratto:
  - *Pantalone-Pericolo:* le parole scelte per definire pantalone risultano più condivise rispetto a quelle scelte per rappresentare pericolo.
  - *Microscopio-Euristica:* in questo caso è euristica ad avere una maggiore condivisione lessicale rispetto a microscopio. Questo può essere spiegato dal fatto che le definizioni sono state date da persone iscritte allo stesso corso di laurea magistrale in informatica e che quindi tendono a utilizzare gli stessi termini.
- Generico/Specifico:
  - *Pantalone-Microscopio:* le definizioni di microscopio hanno meno termini in comune rispetto a pantalone. Ciò può essere dovuto anche alla lunghezza delle definizioni: più una definizione è lunga più è probabile trovare termini che differiscono da un'altra.
  - *Pericolo-Euristica:* poiché pericolo è un termine molto vago ci possono essere molti modi diversi per descriverlo mentre per quanto riguarda euristica meno.

2.1.2 SimSem

Per valutare la similarità semantica si sono utilizzati degli embeddings e si è calcolata la *cosine similarity* a due a due tra i vettori così ottenuti. Infine, come per la similarità lessicale, si è fatta la media.

Table 2.2: Similarità Semantica.

Termine	Similarità Semantica
Pantalone	0.6093
Microscopio	0.4734
Pericolo	0.4783
Euristica	0.0350

Considerazioni:

- Concreto/Astratto:
  - *Pantalone-Pericolo:* entrambi i valori sono relativamente alti, entrambi i termini, pur essendo generici, sono ampliamente conosciuti.
  - *Microscopio-Euristica:* microscopio ha un valore molto più alto di euristica perché tende a essere qualcosa di cui si ha in testa il concetto in modo preciso. Sebbene anche euristica sia specifico e comunque difficile da concepire esattamente data la sua natura astratta.
- Generico/Specifico:
  - *Pantalone-Microscopio:* microscopio essendo meno comune ha registrato una similarità lessicale minore, nonostante fosse probabile il contrario.
  - *Pericolo-Euristica:* a livello semantico le definizioni di euristica hanno un valore incredibilmente basso. Ciò è dovuto al fatto che è un termine difficile da definire con precisione, mentre la sensazione di pericolo è qualcosa di più condivisibile e conosciuto, per cui è più facile condividerne il significato.

2.2 Content to Form

I dizionari consentono la ricerca partendo dal termine per trovare la definizione. Una possibile alternativa è fare una *ricerca onomasiologica*, ossia dalla significato si vuole trovare la parola corrispondente. In quest'esercitazione si utilizzano le definizioni utilizzate nella definizione precedente per testare questa ricerca. Per fare ciò si utilizzano gli iperonimi e il principio del genus. Il mio approccio è quello di prendere come genus la prima parola (escluse stopwords) basandomi sull'intuizione che una buona definizione parta sempre con la categoria generale. Un approccio alternativo potrebbe essere quello di contare la frequenza delle parole e prendere come genus le parole più frequentemente usate nelle definizioni.

Table 2.3: Risultati della ricerca onomasiologica.

Termine	Definizioni Indovinate
Pantalone	35/40
Microscopio	22/39
Pericolo	9/38
Euristica	4/36

Considerazioni: questi risultati indicano che la mia scelta del genus risulta più o meno efficacie per le definizioni concrete (per esempio in pantalone molte definizioni iniziano con "indumento"), mentre se si passa a definizioni astratte la ricerca fallisce più spesso. Ciò potrebbe essere migliorato utilizzando l'approccio basato sulla frequenza descritto sopra.

3 Topic e LLM

3.1 Topic Modelling

Per questa esercitazione si è scelto di usare il dataset `zou-lab/MedCaseReasoning` di *Hugging Face* che contiene 14.489 righe di dati contenenti articoli medici. Osservando il grafico prodotto (fig. 3.1) si può notare una coerenza semantica tra le keywords:

- Il topic 0 sembra riferirsi a malattie cardiache (tachycardia, cardiomyopathy,heart).
- Il topic 1 a malattie del sistema nervoso (encephalopathy, meningitis, encephalitis).
- etc.

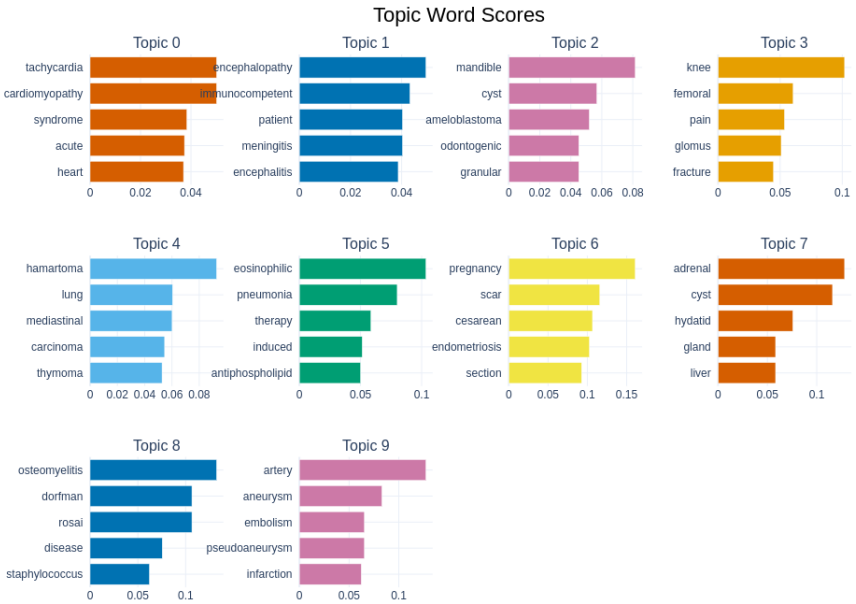

Figure 3.1: Grafico a barre raffigurante i top topics.

3.2 LLM Prompting

In quest'ultima esercitazione si sono esplorate varie strategie di prompting e dei parametri dei modelli LLM.

3.2.1 Label the Topic

Il task consiste nell'assegnare un'etichetta ai topics prodotti nell'esercitazione precedente:

- *Zero-shot:* invece che assegnare una label ai topic assume che tutti i concetti appartengano a un unico paziente.
- *One-shot:* con un esempio il task diventa più preciso riuscendo a riassumere le caratteristiche comuni in un'unica etichetta.

3.2.2 Guess from the Definitions

Per tentare di far indovinare le parole dalla definizione si è utilizzato un approccio simile a quello usato per i topics:

- *Zero-shot:* il modello fraintende il task assegnato e invece di restituire le parole restituisce una nuova definizione.
- *One-shot:* aggiungendo un esempio sull'output atteso i risultati migliorano di molto. Il modello riesce a individuare correttamente le parole *microscopio* e *pericolo*, mentre è troppo generico su *pantalone* (tenta con indumento) ed *euristica* (linee guida).
- *One-shot with suggestions:* aggiungendo al prompt precedente dei "suggerimenti", ossia le definizioni prese dal vocabolario della Treccani, si nota un miglioramento per la parola *pantalone* che adesso viene riconosciuta correttamente più spesso. Per quanto riguarda *euristica* viene restituito metodologia.