Relazione
Esercitazione 2

Content2Form

Introduzione

La seguente esercitazione ha affrontato il problema della ricerca onomasiologica. Questo task consiste nell'individuare un concetto a partire dalla sua definizione. Come verrà evidenziato in seguito, anche la ricerca onomasiologica rappresenta un problema difficile da affrontare.

Nello specifico in questa esercitazione si prova a risalire al synset partendo dalle definizioni della risorsa *definizioni.csv* già utilizzata e descritta nell'esercitazione precedente (Defs).

Ai fini della valutazione, sono stati individuati ed implementati due approcci

- genus (targeted search)
  sono stati individuati i termini più frequenti nelle definizioni, di questi sono stati presi i primi `N_GENUS_WORDS (6)` che saranno appunto i genus, in questo modo è possibile restringere la ricerca in Wordnet, confrontando le definizioni con un sottoinsieme di definizioni specifico, passando da migliaia di definizioni ad un centinaio, in modo da rendere la ricerca più efficiente

- global (complete search)
  è stata condotta una ricerca confrontando la similarità di ogni definizione (per ogni concetto) con le gloss di WordNet.

NOTE:

- Per entrambe le ricerche vengono restituiti i 5 migliori synsets individuati per ogni concetto

Struttura del codice ed implementazione

Per lo svolgimento di questa esercitazione il codice è presente nei file

- **`main.py`**

  che si occupa di richiamare i metodi che implementano i due approcci prima indicati e stampare i risultati

- **`utils.py`**

  che implementa i seguenti metodi

  - `find_synset_by_gloss(gloss)`

    restituisce il synset data la glossa in input

  - `defs_dict()`

    costruisce un dizionario con `chiave = concetto` e `valore = lista delle definizioni del concetto`

  - `similarity(sent_1, sent_2)`

    metodo che misura la similarità tra due frasi date in input tramite la cosine similarity

  - `best_def_similariry(list_sent)`

    calcola la similarità (appoggiandosi al metodo `similarity`) tra della lista di definizioni annotate e le gloss di WN. Ritorna una lista con le prime N_SYNSETS (5) migliori definizioni con relativo score e synset associato.

  - `onomasiological_by_word_freq()`

    metodo che implementa il primo approccio per la ricerca onomasiologica:
    - viene definito il genus andando a reperire le parole più frequenti dalle definizioni annotate (tramite il metodo `parse_input`, descritto nella precedente esercitazione.
    - Le frequenze delle parole del genus vengono successivamente normalizzate al fine di ottenere uno score per ognuna.
    - Per ogni concetto, viene eseguito un ciclo su tutte le gloss di WN andando ad assegnare uno score ogni volta in cui in una glossa è presente almeno una parola del genus
    - Una volta terminato, lo score viene normalizzato e viene restituita una lista contenente i primi candidati `N_SYNSETS (5)`

  - `onomasiological_by_sentences_sim()`

    metodo che implementa il secondo approccio per la ricerca onomasiologica:
    - viene effettuata una ricerca completa
    - le definizioni vengono pre-processate andando a rimuovere quelle vuote

- per ogni definizione di ogni concetto viene misurata la similarità, tramite la cosine similarity, con le gloss di WN, mantenendo in memoria solo la gloss che ha ottenuto similarità maggiore.
- una volta terminata la ricerca per ogni concetto viene restituita una lista con il miglior risultato o i migliori nel caso lo score non sia pari a 100

Risultati

Di seguito i risultati ottenuti

- metodo genus (targeted search)

Concept	Genus words	WordNet gloss	Correct synset	Best synsets found
Emotion	feeling, 0.38 human, 0.25 feel, 0.25 something, 0.22 state', 0.12 living', 0.12	*any strong feeling*	Synset('emotion.n.01')	(Synset('bloodless.s.04'), 0.14) (Synset('cold.s.09'), 0.14) (Synset('pitiless.s.02'), 0.14) (Synset('dead.s.06'), 0.14) (Synset('pathetic_fallacy.n.01'), 0.14)
Person	human, 0.91 person, 0.19 living, 0.12 individual, 0.09 certain, 0.09 ability, 0.09	*a human being*	Synset('person.n.01')	Synset('reincarnation.n.03'), 0.20 Synset('depersonalization.n.03'), 0.19 Synset('speaking_trumpet.n.01'), 0.18 Synset('body_temperature.n.01'), 0.18 Synset('theophany.n.01'), 0.18
Revenge	someone, 0.44 anger, 0.25 feeling, 0.22 action, 0.19 emotion, 0.19 reaction, 0.19	*action taken in return for an injury or offense*	Synset('revenge.v.01')	Synset('feel.v.05'), 0.17 Synset('lightning_rod.n.01'), 0.13 Synset('argonaut.n.01'), 0.11 Synset('crazy.n.01'), 0.11 Synset('decoy.n.01'), 0.11
Brick	used, 0.75 object, 0.5 material, 0.5 construction 0.5 build, 0.41 building, 0.31	*rectangular block of clay baked by the sun or in a kiln; used as a building or paving material*	Synset('brick.n.01')	Synset('brick.n.01'), 0.32 Synset('building_material.n.01'), 0.32 Synset('cement.n.02'), 0.32 Synset('fieldstone.n.01'), 0.32 Synset('flooring.n.02'), 0.32

- metodo global (complete search)

Concept	Best defs found	WordNet gloss	Score	Best synsets found	Correct synset
Emotion	*a strong feeling*	*any strong feeling*	100	Synset('emotion.n.01')	Synset('emotion.n.01')
Person	*human being*	*a human being*	100	Synset('person.n.01')	Synset('person.n.01')
Revenge	*act of doing something to someone because of anger*	*the act of losing someone or something*	0.75	Synset('loss.n.03')	Synset('revenge.v.01')
	the act of damaging someone as a reaction	*the act of damaging something or someone*	0.75	Synset('damage.n.03')	
	negative emotion	*a negative*	0.70	Synset('no.n.01')	
	desire	*having or feeling no desire*	0.70	Synset('undesirous.a.01')	
	mood	*in a bad mood*	0.70	Synset('cantankerously.r.01')	
Brick	*block of some material, used in construction*	*a block of material used in construction work*	0.8	Synset('building_block.n.02')	Synset('brick.n.01')
	block of material used for building construction		0.8		
	material used for construction		0.77		
	part of build	*for the most part*	0.70	Synset('chiefly.r.01')	
	object used to build something	*what something is used for*	0.70	Synset('function.n.02')	

Conclusioni

Come si può notare dai seguenti risultati, la ricerca onomasiologica risulta essere un task non semplice. Fra i possibili risultati, una certa attenzione merita il concetto *brick* il quale per entrambi i metodi ha portato a ottimi risultati. Per quanto riguarda gli altri concetti molto spesso i synsets trovati, seppur non corretti, si avvicinano, almeno da un punto di vista del campo semantico restituito

Nello specifico per il metodo del genus il concetto *brick* ha prodotto risultati eccellenti restituendo in prima posizione (score maggiore) il synset corretto, ma anche i suoi risultati successivi sono semanticamente legati, mentre per gli altri concetti non si è arrivati al synset corretto.

Per il metodo global i concetti *emotion* e *person* hanno riportato una precisione del 100% nell'identificazione del synset, meno precisione invece per brick, mentre per *revenge* non è stato identificato

Possibili motivazioni a seguito dei risultati ottenuti

- i genus: come possiamo vedere i genus di riferimento hanno poco a che vedere con i termini, come ad esempio *"someone"* per *revenge*. Questo è portato dal dataset in input. Si potrebbe ripulire il dataset rimuovendo le parole che non hanno a che vedere con il termine originale, ma così facendo si andrebbe a compromettere l'esercizio e si renderebbe questo approccio poco scalabile su altre basi di dati.

- funzione di similarità: Sarebbe utile far girare l'algoritmo utilizzando altre misure di similarità oltre la *cosine similarity* per confrontare i risultati ottenuti decidere anche di utilizzare altre funzioni di similarità oltre a quella basata sulla comparazione dei termini più frequenti.

Per concludere, anche se i risultati non sono ottimi, non sorprende che il conetto ad aver avuto successo sia stato appunto *brick* siccome, almeno in teoria, è il termine più facile a cui dare una definizione, essendo concreto e specifico.

Revenge, seppur astratto e specifico, non ha restituito risultati soddisfacenti, nonostante *emotion* (astratto e generico), per il secondo metodo, ha restituito il synset corretto.